

# Adversarial Attacks in the Audio Domain

CS349 Machine Learning  
Northwestern University  
12.1.21

Patrick O'Reilly



[github.com/oreillyp/adv\\_audio\\_intro](https://github.com/oreillyp/adv_audio_intro)

# Adversarial Examples Fool Neural Networks

---

# Adversarial Examples Fool Neural Networks

---

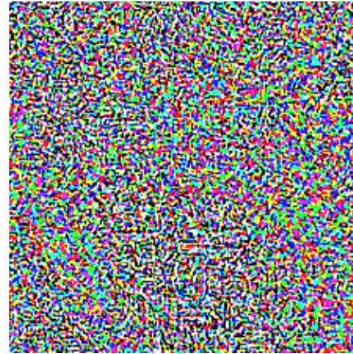


$x$

“panda”

57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”

8.2% confidence

=



$x +$

$\epsilon \text{sign}(\nabla_x J(\theta, x, y))$

“gibbon”

99.3 % confidence

(Goodfellow et al. 2014)

# Adversarial Examples Fool Neural Networks

---

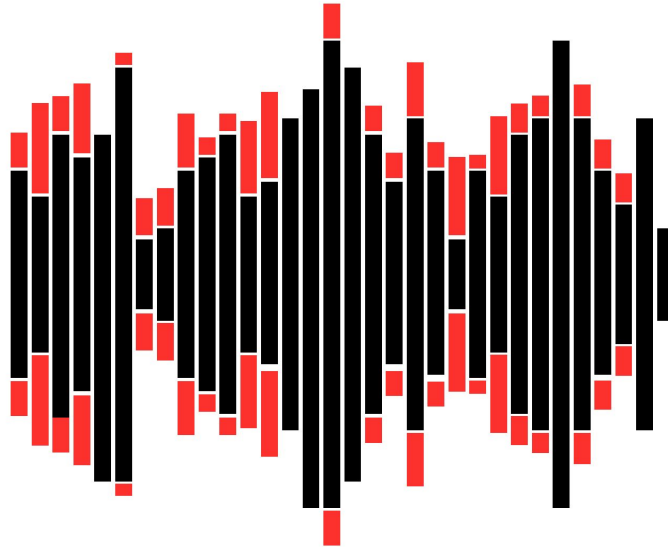
$x$



# Adversarial Examples Fool Neural Networks

---

$$x + \delta$$



# Neural Networks Power Voice Interfaces

Voice-based machine-learning systems for authentication and control are common in products such as mobile devices, vehicles, and household appliances.



# What Systems Might Attackers Target?

---

# What Systems Might Attackers Target?

---

Recognize “Hey  
Alexa,” “OK  
Google,” “stop,”  
“go,” ...

*Wake-word detection,  
speech command  
recognition*

Verify a  
speaker’s  
identity (against  
enrolled profile)

*Automatic speaker  
verification, speaker  
recognition*

Transcribe all  
incoming speech

*Automatic speech  
recognition*



**...Has Anyone Looked Into This?**

---

# ...Has Anyone Looked Into This?

Google Scholar

audio adversarial examples

Articles About 48,200 results (0.06 sec)

**Any time**  
Since 2021  
Since 2020  
Since 2017  
Custom range...

**Sort by relevance**  
Sort by date

**Any type**  
Review articles

include patents  
 include citations

Create alert

**Audio adversarial examples: Targeted attacks on speech-to-text**  
[N Carlini, D Wagner](#) - 2018 IEEE Security and Privacy ..., 2018 - [ieeexplore.ieee.org](#)  
We construct targeted audio adversarial examples on automatic speech recognition. Given any audio waveform, we can produce another that is over 99.9% similar, but transcribes as any phrase we choose (recognizing up to 50 characters per second of audio). We apply our ...  
☆ Save 📄 Cite Cited by 714 Related articles All 11 versions

**Imperceptible, robust, and targeted adversarial examples for automatic speech recognition**  
[Y Qin, N Carlini, G Cottrell](#)... - ... on machine learning, 2019 - [proceedings.mlr.press](#)  
... **adversarial examples**, we depart from the common lp distance measure widely used for **adversarial example** research. Instead, we make use of the psychoacoustic principle of auditory masking, and only add the **adversarial** perturbation to regions of the **audio** where it will not ...  
☆ Save 📄 Cite Cited by 188 Related articles All 11 versions

**Characterizing audio adversarial examples using temporal dependency**  
[Z Yang, B Li, PY Chen, D Song](#) - arXiv preprint arXiv:1809.10875, 2018 - [arxiv.org](#)  
Recent studies have highlighted **adversarial examples** as a ubiquitous threat to different neural network models and many downstream applications. Nonetheless, as unique data properties have inspired distinct and powerful learning principles, this paper aims to explore ...  
☆ Save 📄 Cite Cited by 75 Related articles All 6 versions

# ...Has Anyone Looked Into This?

The image shows a Google Scholar search interface. The search bar contains the text "audio adversarial examples". Below the search bar, there are two search results. The first result is titled "Audio adversarial examples: Targeted attacks on speech-to-text" by N. Carlini and D. Wagner, published in 2018. The second result is titled "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition" by Y. Qin, N. Carlini, and G. Cottrell, published in 2019. A third result is partially visible, titled "Characterizing audio adversarial examples using temporal dependency" by Z. Yang, B. Li, P. Y. Chen, and D. Song, published in 2018. A large red text overlay is placed over the first result, stating "~20-30 relevant attacks published, most since 2018".

Google Scholar

audio adversarial examples

Articles

About 48,200 results (0.06 sec)

Any time

Since 2021

Since 2020

Since 2017

Custom range...

Sort by relevance

Sort by date

Any type

Review articles

include patents

include citations

Create alert

**~20-30 relevant attacks published, most since 2018**

**Audio adversarial examples: Targeted attacks on speech-to-text**  
N. Carlini, D. Wagner - 2018 IEEE S&P: Security and Privacy, 2018 - ieeexplore.ieee.org  
... adversarial examples, we depart from the common lp distance measure widely used for adversarial example research. Instead, we make use of the psychoacoustic principle of auditory masking, and only add the adversarial perturbation to regions of the audio where it will not ...  
☆ Save Cite Cited by 188 Related articles All 11 versions

**Imperceptible, robust, and targeted adversarial examples for automatic speech recognition**  
Y. Qin, N. Carlini, G. Cottrell... - ... on machine learning, 2019 - proceedings.mlr.press  
... adversarial examples, we depart from the common lp distance measure widely used for adversarial example research. Instead, we make use of the psychoacoustic principle of auditory masking, and only add the adversarial perturbation to regions of the audio where it will not ...  
☆ Save Cite Cited by 188 Related articles All 11 versions

**Characterizing audio adversarial examples using temporal dependency**  
Z. Yang, B. Li, P. Y. Chen, D. Song - arXiv preprint arXiv:1809.10875, 2018 - arxiv.org  
Recent studies have highlighted adversarial examples as a ubiquitous threat to different neural network models and many downstream applications. Nonetheless, as unique data properties have inspired distinct and powerful learning principles, this paper aims to explore ...  
☆ Save Cite Cited by 75 Related articles All 6 versions

# ...Has Anyone Looked Into This?

The image shows a Google Scholar search interface. The search bar contains the text "audio adversarial examples". Below the search bar, there are two search results. The first result is titled "Audio adversarial examples: Targeted attacks on speech-to-text" by N. Carlini and D. Wagner, published in 2018. The second result is titled "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition" by Y. Qin, N. Carlini, and G. Cottrell, published in 2019. Overlaid on the search results is large red text that reads: "~20-30 relevant attacks published, most since 2018" and "A similar number of defenses have been proposed".

Google Scholar

audio adversarial examples

Articles

About 48,200 results (0.06 sec)

Any time

Since 2021

Since 2020

Since 2017

Custom range...

Sort by relevance

Sort by date

Any type

Review articles

include patents

include citations

Create alert

**~20-30 relevant attacks published, most since 2018**

**A similar number of defenses have been proposed**

**Audio adversarial examples: Targeted attacks on speech-to-text**  
N. Carlini, D. Wagner - 2018 IEEE Symposium on Security and Privacy, 2018 - ieeeexplore.ieee.org  
We have recently demonstrated targeted attacks on automatic speech recognition. Given any audio waveform, we can produce another that is over 99.9% similar, but transcribes as any phrase we choose (recognizing approximately 160 characters per second of audio). We apply our ...  
☆ Save Cite Cited by 188 Related articles All 11 versions

**Imperceptible, robust, and targeted adversarial examples for automatic speech recognition**  
Y. Qin, N. Carlini, G. Cottrell... - ... on machine learning, 2019 - proceedings.mlr.press  
... **adversarial examples**, we depart from the common lp distance measure widely used for **adversarial examples** search. Instead, we make use of the psychoacoustic principle of auditory masking, which allows us to perturb the audio waveform in regions where it will not ...  
☆ Save Cite Cited by 188 Related articles All 11 versions

**Characterizing audio adversarial examples using temporal dependency**  
Z. Yang, B. Li, P. Y. Chen, D. Song - arXiv preprint arXiv:1809.10875, 2018 - arxiv.org  
Recent studies have demonstrated **adversarial examples** as a ubiquitous threat to different neural network models and their downstream applications. Nonetheless, as unique data properties have inspired distinct and powerful learning principles, this paper aims to explore ...  
☆ Save Cite Cited by 75 Related articles All 6 versions

# What Systems Might Attackers Target?

---

Recognize “Hey  
Alexa,” “OK  
Google,” “stop,”  
“go,” ...

*Wake-word detection,  
speech command  
recognition*

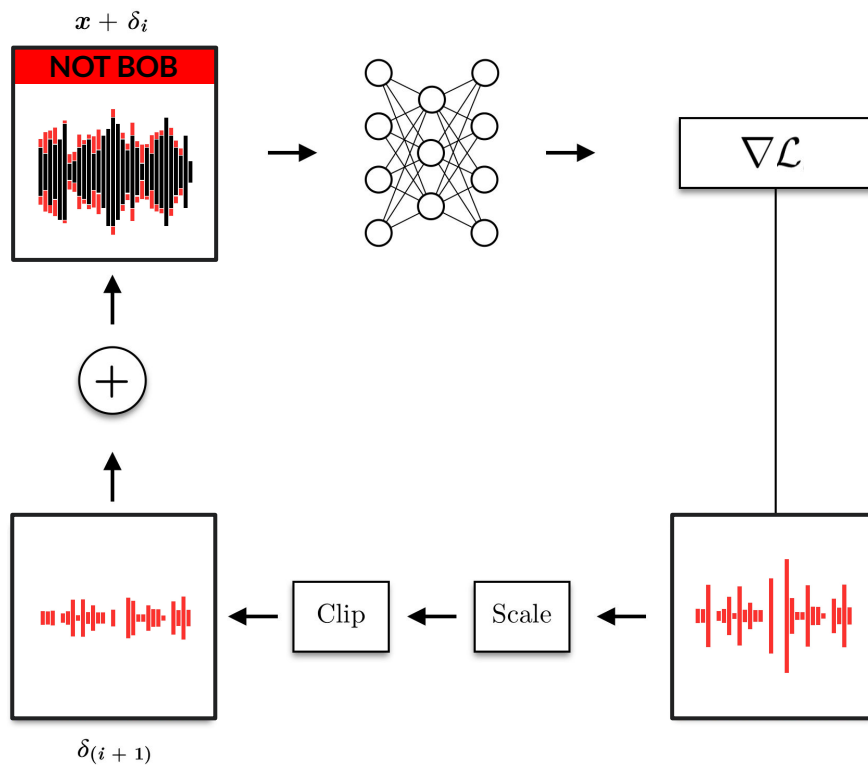
Verify a  
speaker’s  
identity (against  
enrolled profile)

*Automatic speaker  
verification, speaker  
recognition*

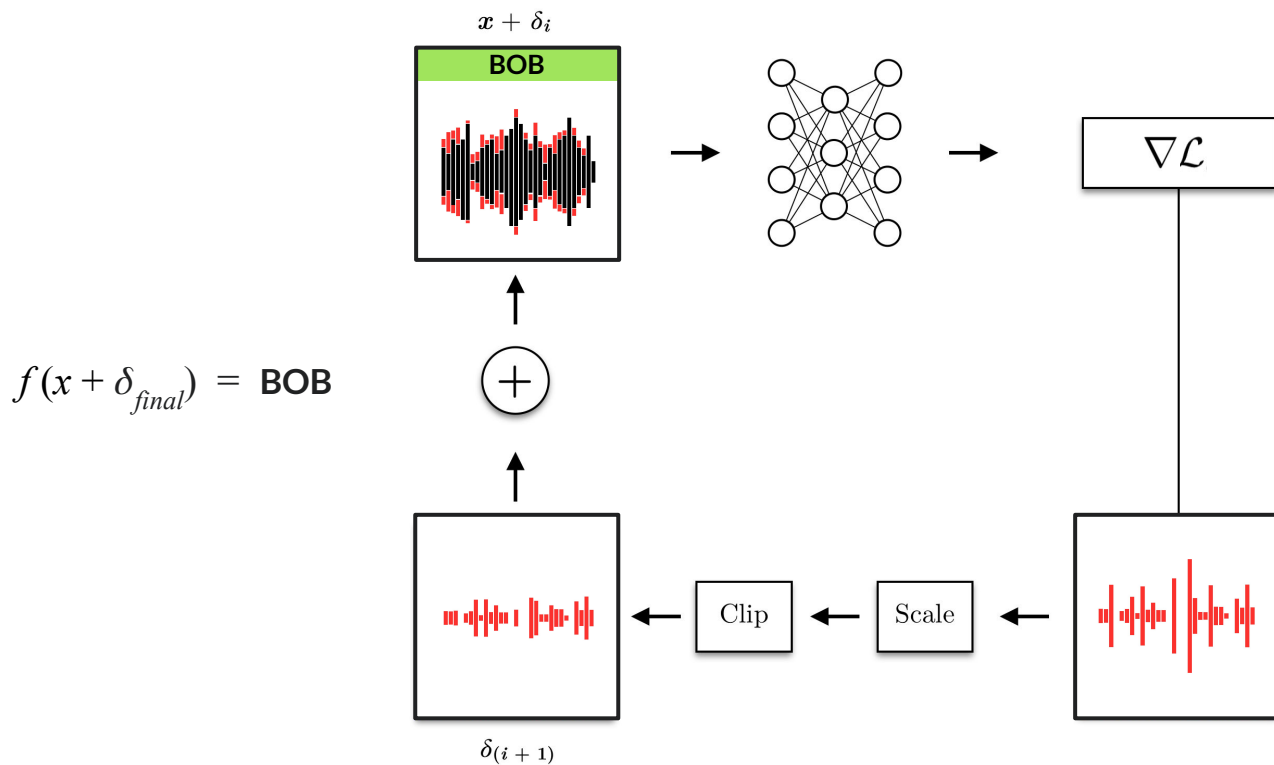
Transcribe all  
incoming speech

*Automatic speech  
recognition*

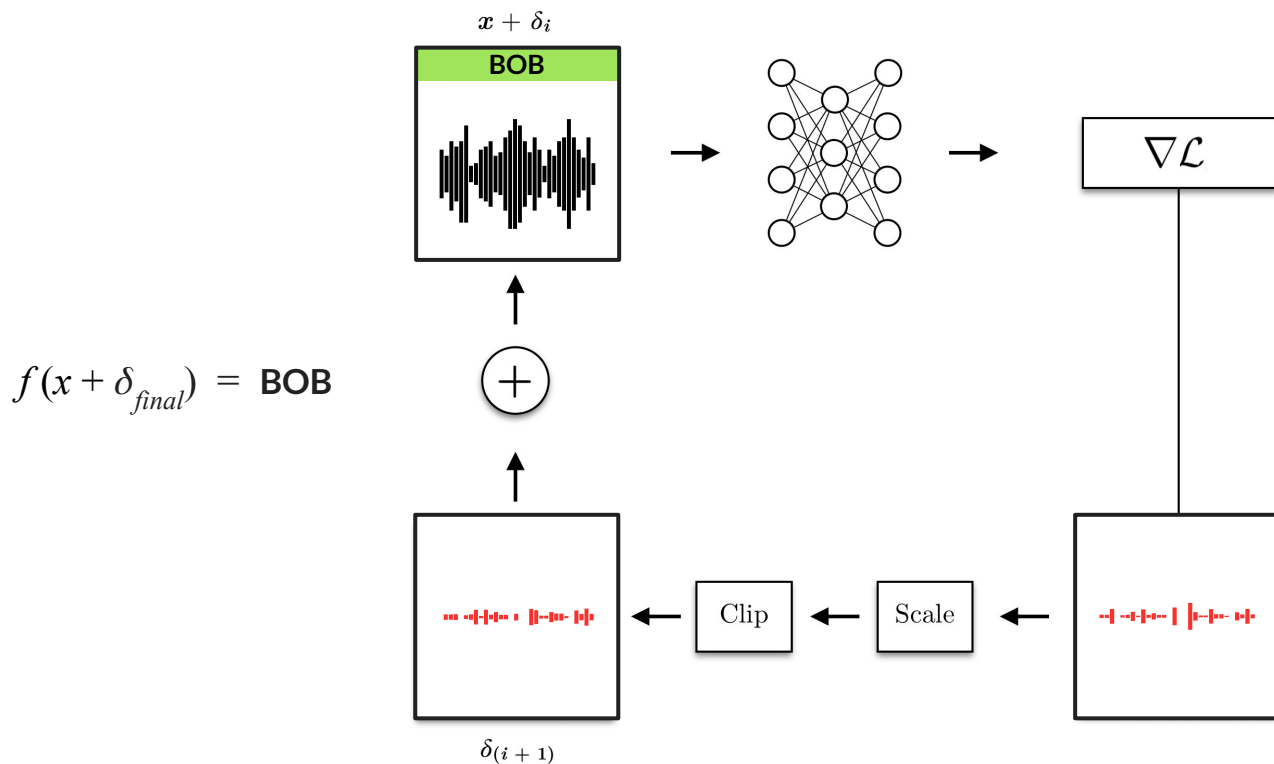
# How Do We Make Adversarial Examples?



# How Do We Make Adversarial Examples?



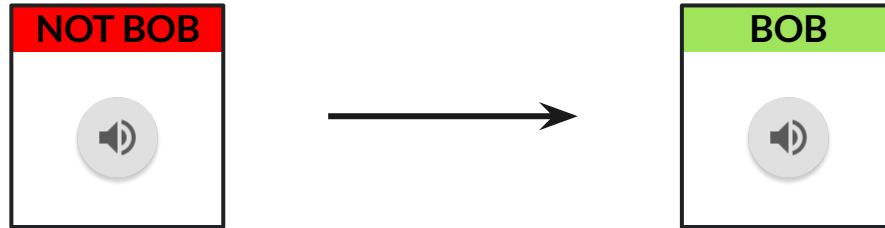
# How Do We Make Adversarial Examples?





# How Should We Attack?

---



# Effective and Inconspicuous Over-the-Air Adversarial Examples with Adaptive Filtering

Patrick O'Reilly<sup>1</sup>, Pranjal Awasthi<sup>2</sup>, Aravindan Vijayaraghavan<sup>1</sup>, Bryan Pardo<sup>1</sup>

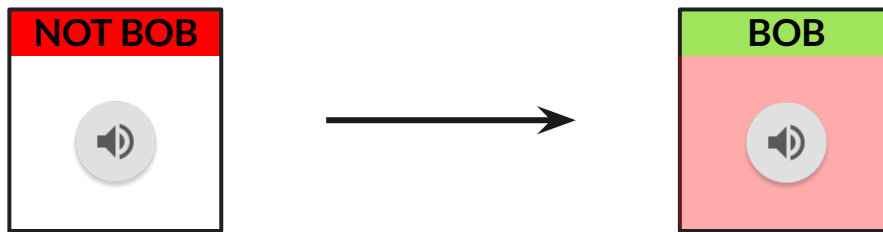
*Submitted to ICASSP '22*



1. Northwestern University
2. Google Research

[interactiveaudiolab.github.io/project/audio-adversarial-examples.html](https://interactiveaudiolab.github.io/project/audio-adversarial-examples.html)

# How Should We Attack?



"Generic"

Approach

**image-domain** (sample-wise additive noise)

Perceptual Regularization

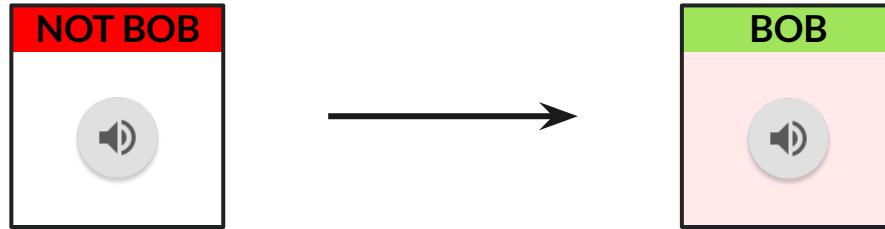
**simple** ( $L_2$  penalty)

Perceptual Quality

**poor** (perturbation is obvious)

# How Should We Attack?

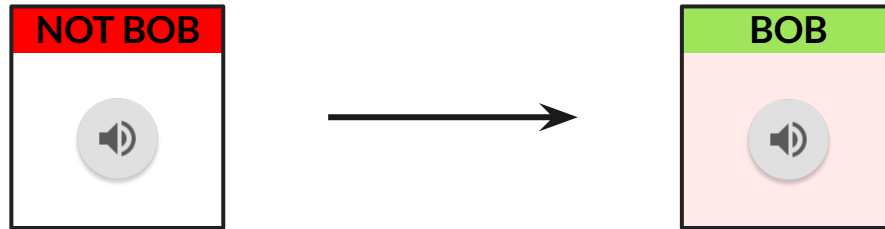
\*Qin et al. (2019),  
Szurley & Kolter (2019),  
Dörr et al. (2020),  
Wang et al. (2020)



	"Generic"	Qin et al.*
Approach	<b>image-domain</b> (sample-wise additive noise)	<b>image-domain</b> (sample-wise additive noise)
Perceptual Regularization	<b>simple</b> ( $L_2$ penalty)	<b>complex</b> (frequency masking loss)
Perceptual Quality	<b>poor</b> (perturbation is obvious)	<b>good</b> (perturbation is subtle)

# How Should We Attack?

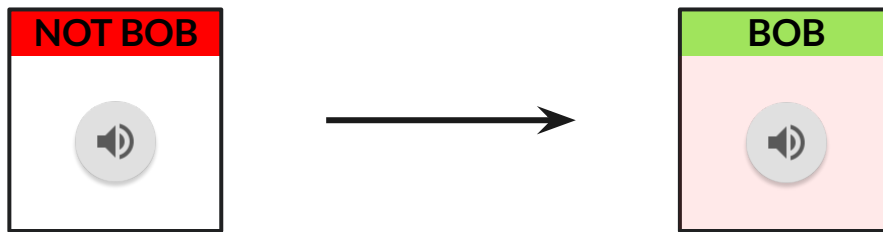
\*Qin et al. (2019),  
Szurley & Kolter (2019),  
Dörr et al. (2020),  
Wang et al. (2020)



	"Generic"	Qin et al.*	Proposed
Approach	<b>image-domain</b> (sample-wise additive noise)	<b>image-domain</b> (sample-wise additive noise)	<b>audio-domain</b> (adaptive filtering)
Perceptual Regularization	<b>simple</b> ( $L_2$ penalty)	<b>complex</b> (frequency masking loss)	<b>simple</b> ( $L_2$ penalty)
Perceptual Quality	<b>poor</b> (perturbation is obvious)	<b>good</b> (perturbation is subtle)	<b>good</b> (perturbation is subtle)

# How Should We Attack?

\*Qin et al. (2019),  
Szurley & Kolter (2019),  
Dörr et al. (2020),  
Wang et al. (2020)

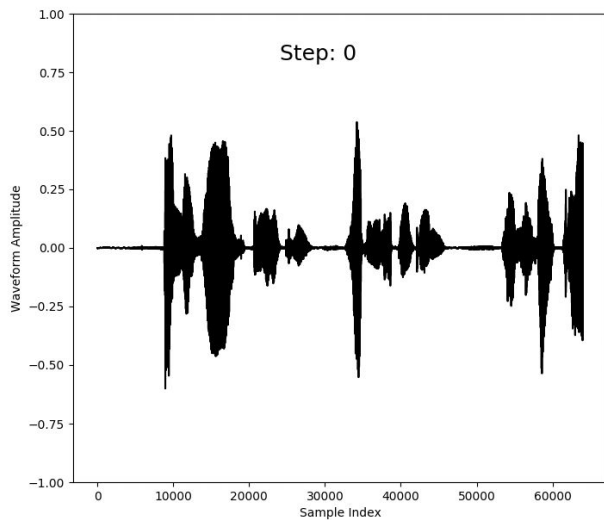


	"Generic"	Qin et al.*	Proposed
Approach	✗	✗	✓
Perceptual Regularization	✓	✗	✓
Perceptual Quality	✗	✓	✓

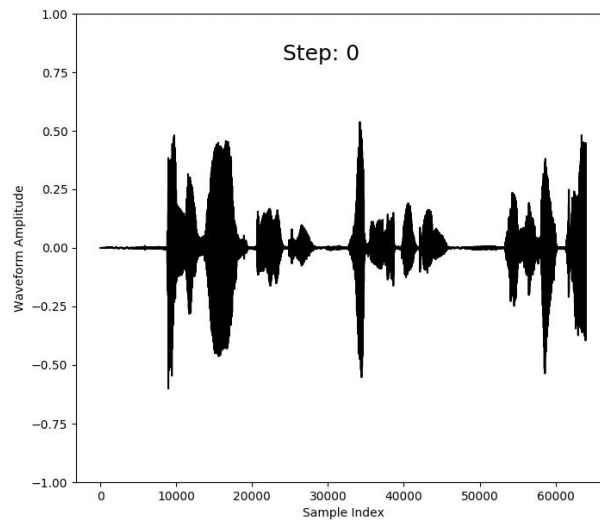
# How Should We Attack?

---

Qin et al.\*



Proposed



# Adversarial Examples Fool Neural Networks

---

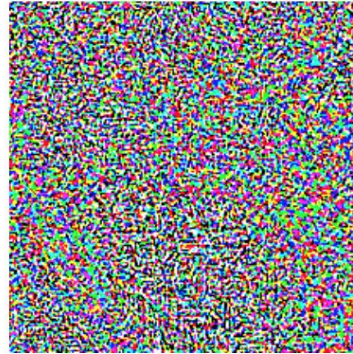


$x$

“panda”

57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”

8.2% confidence

=



$x +$

$\epsilon \text{sign}(\nabla_x J(\theta, x, y))$

“gibbon”

99.3 % confidence

(Goodfellow et al. 2014)



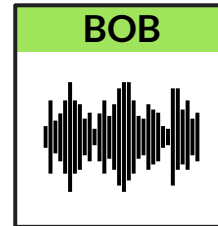
# Let's Attack a Voice Interface

---

# Let's Attack a Voice Interface: Pick a Task

---

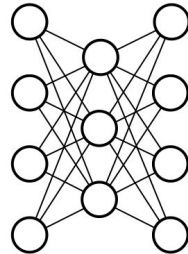
**Speaker Verification:** confirm a speaker's claimed identity (against enrolled profile)



# Let's Attack a Voice Interface: Pick a Task

---

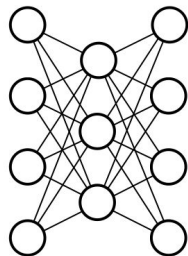
We want a **large** and **accurate** model, as in many applications (e.g. mobile banking) speaker verification models are deployed in the cloud rather than on-device.



# Let's Attack a Voice Interface: Pick a Task

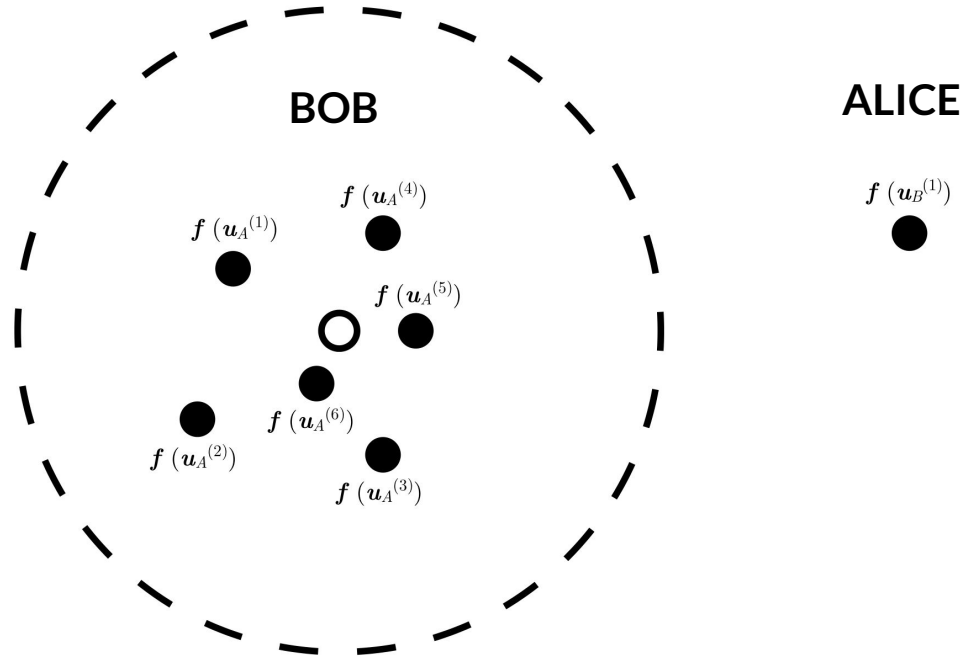
---

Specifically, we'll use the **ResNetSE34V2** model proposed by Heo et al. (2020), available at [https://github.com/clovaai/voxceleb\\_trainer](https://github.com/clovaai/voxceleb_trainer)



# Let's Attack a Voice Interface: Pick an Objective

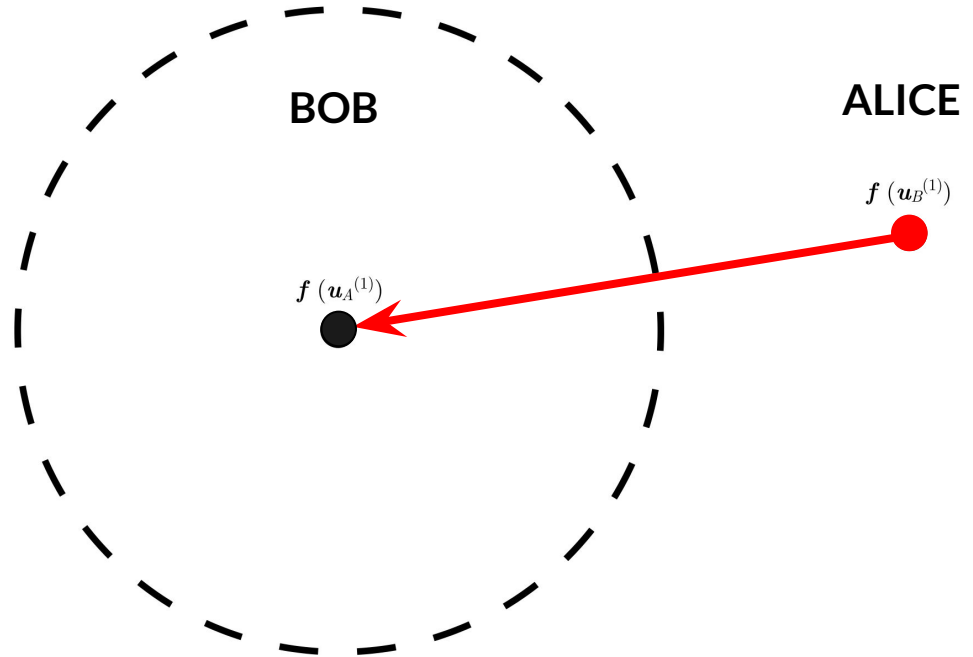
---



# Let's Attack a Voice Interface: Pick an Objective

---

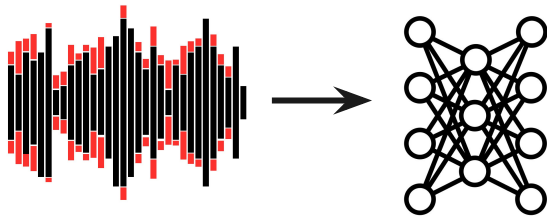
Following Zhang et al. (2021), for the sake of simplicity we will attempt to spoof the embedding of a single utterance.



# Let's Attack a Voice Interface: Pick a Setting

---

**Over-the-line setting:** the attack audio can be fed directly to the victim model over a purely digital channel.

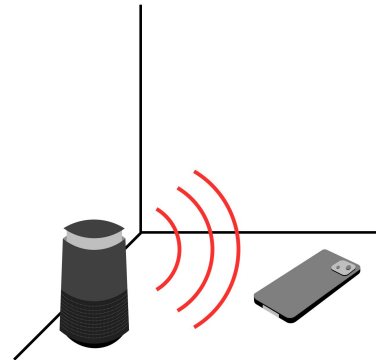
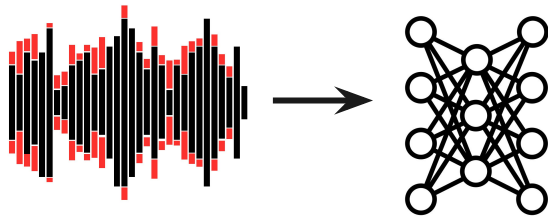


# Let's Attack a Voice Interface: Pick a Setting

---

**Over-the-line setting:** the attack audio can be fed directly to the victim model over a purely digital channel.

**Over-the-air setting:** malicious audio is played through a speaker and received by a microphone before entering the victim model.

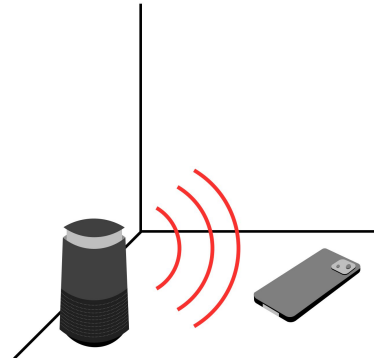
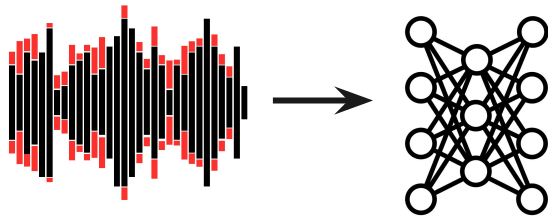




# Let's Attack a Voice Interface: Pick a Setting

**Over-the-line setting:** the attack audio can be fed directly to the victim model over a purely digital channel.

**Over-the-air setting:** malicious audio is played through a speaker and received by a microphone before entering the victim model.



# Let's Attack a Voice Interface: Pick a Setting

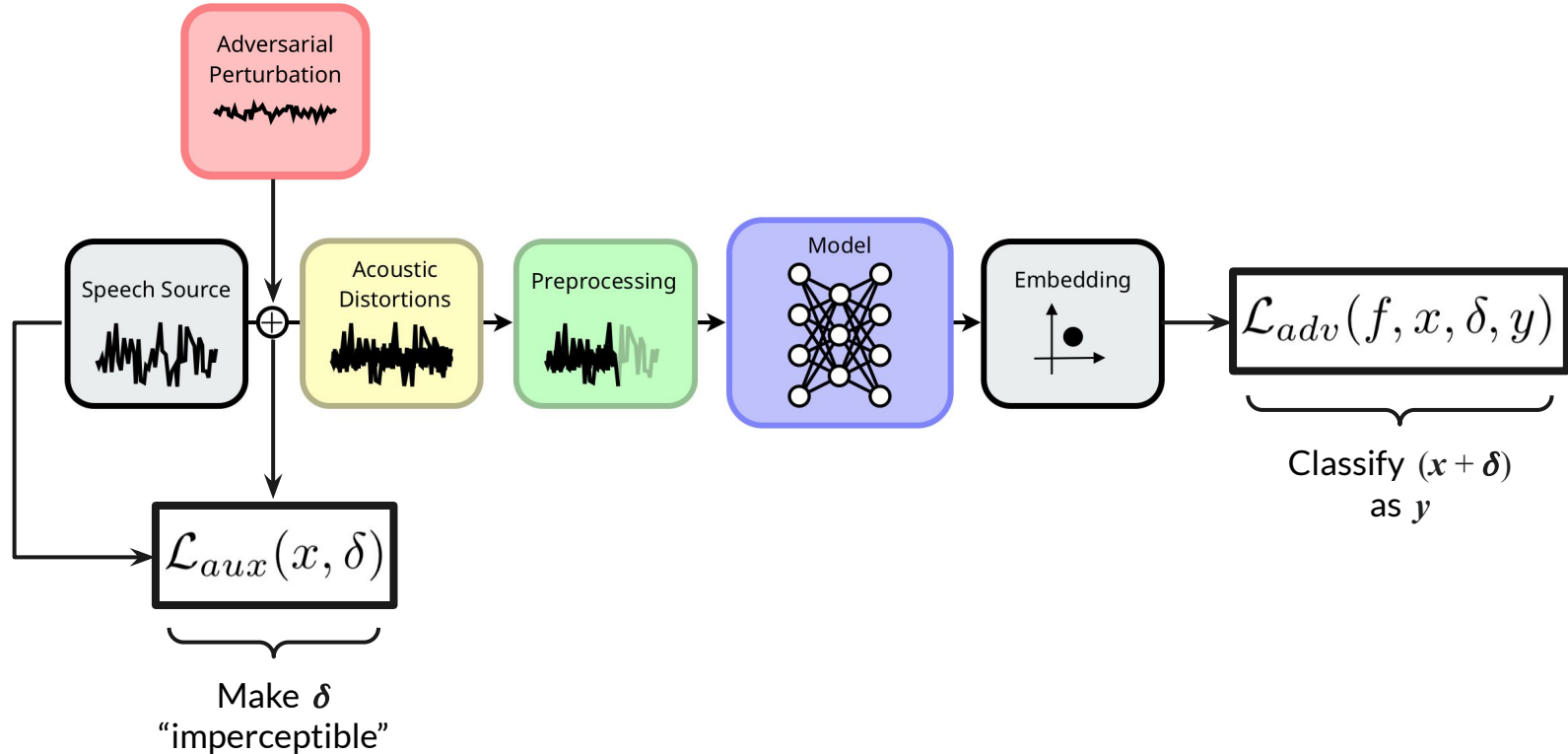
---

**Over-the-line setting:** the attack audio can be fed directly to the victim model over a purely digital channel.

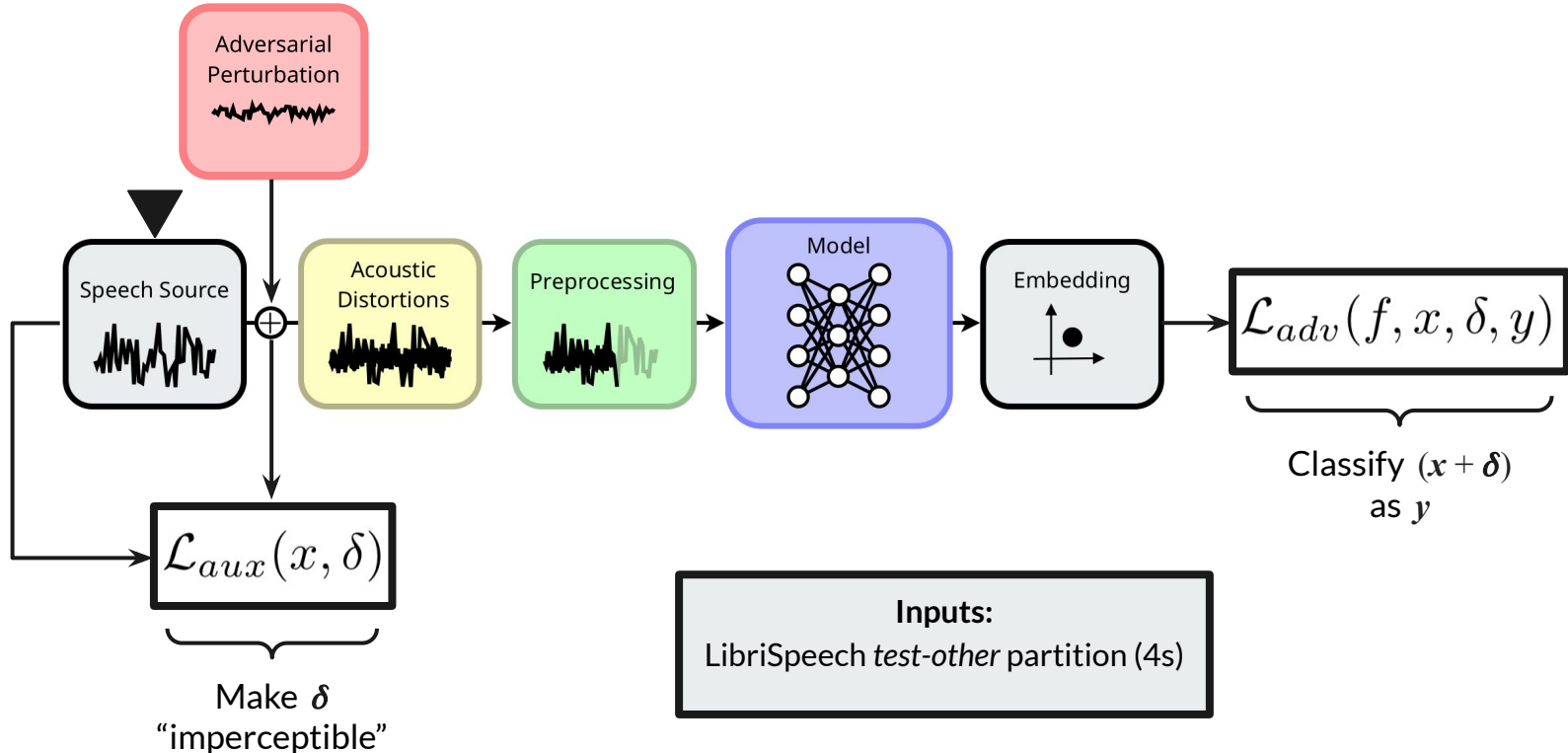
**Over-the-air setting:** malicious audio is played through a speaker and received by a microphone before entering the victim model.



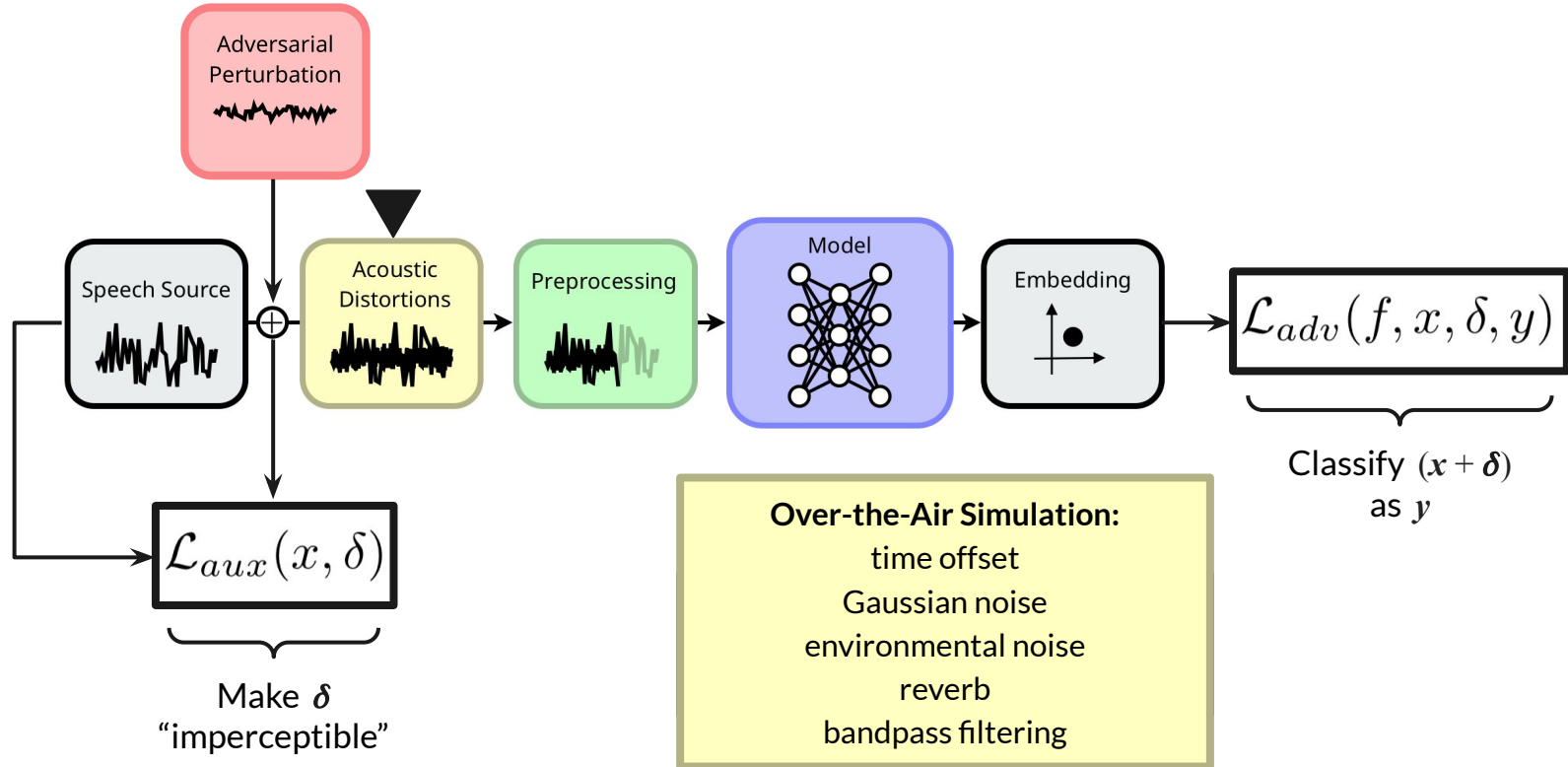
# Let's Attack a Voice Interface: System Design



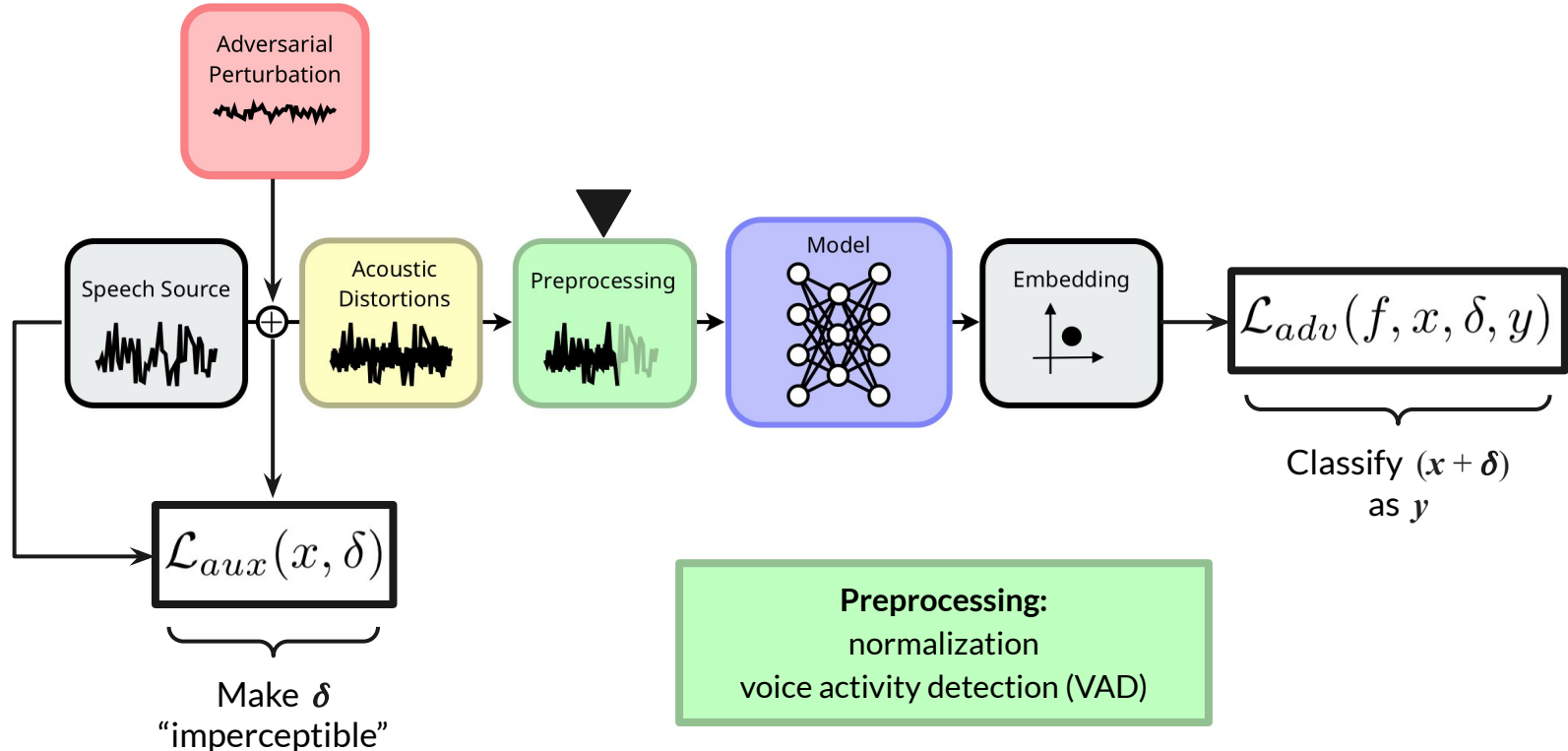
# Let's Attack a Voice Interface: System Design



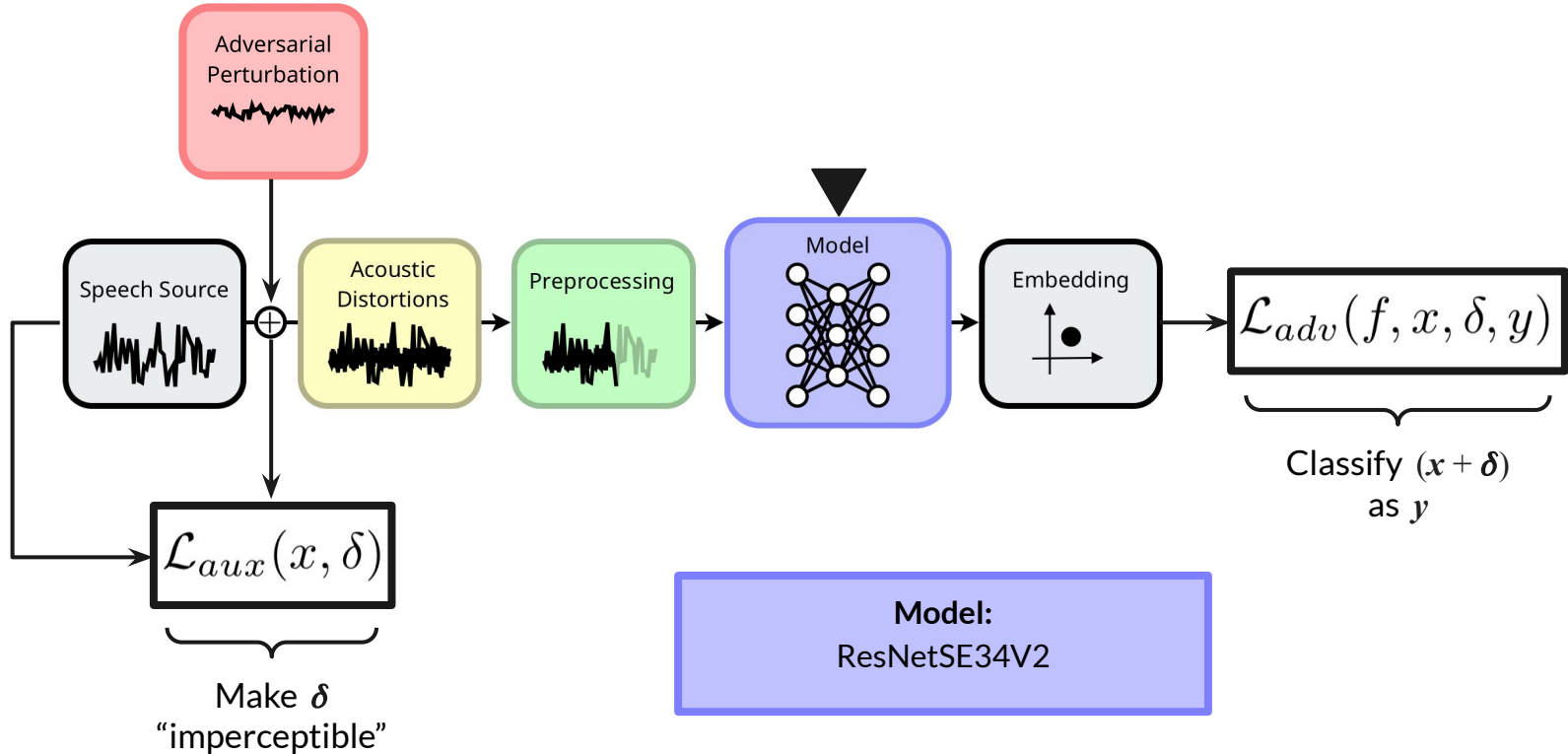
# Let's Attack a Voice Interface: System Design



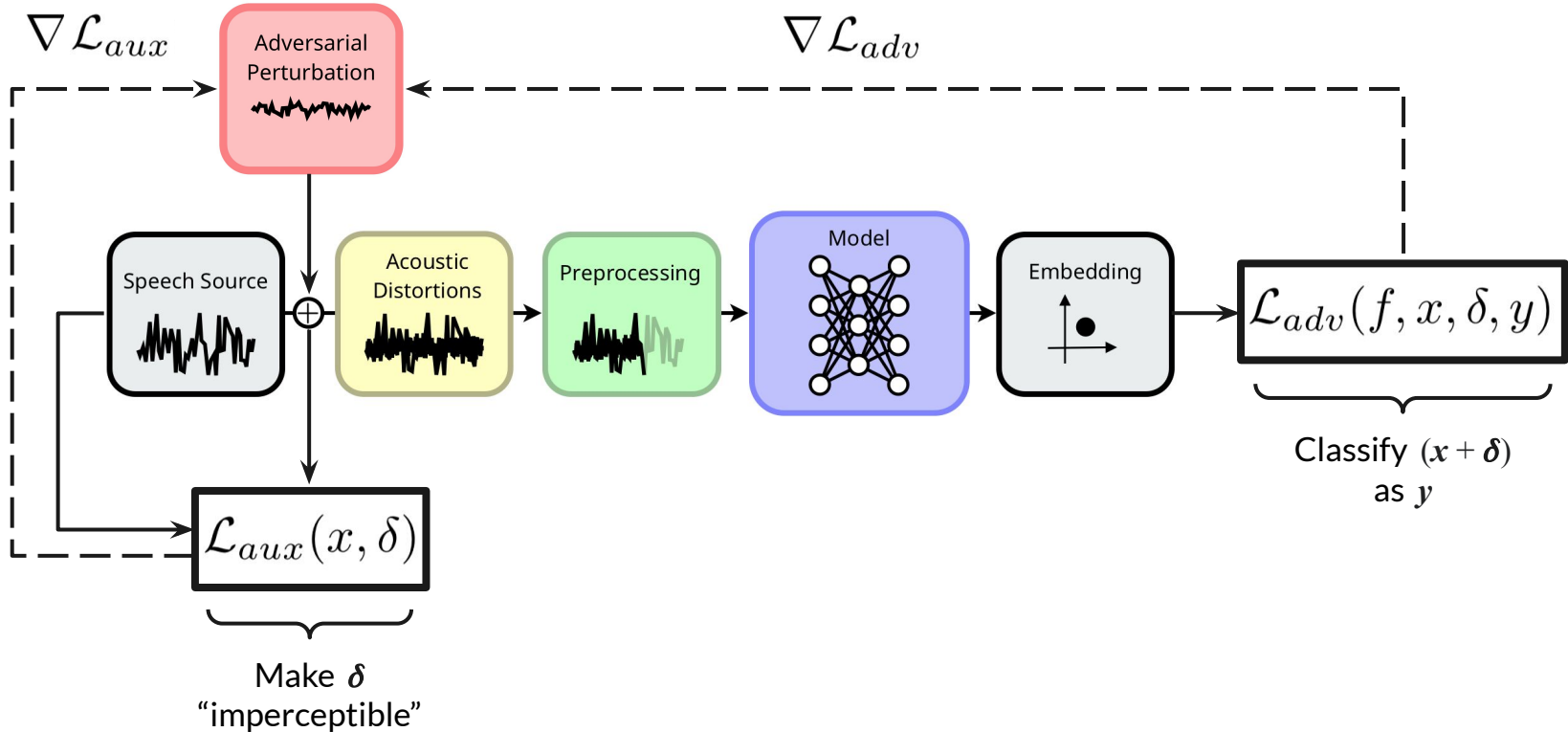
# Let's Attack a Voice Interface: System Design



# Let's Attack a Voice Interface: System Design

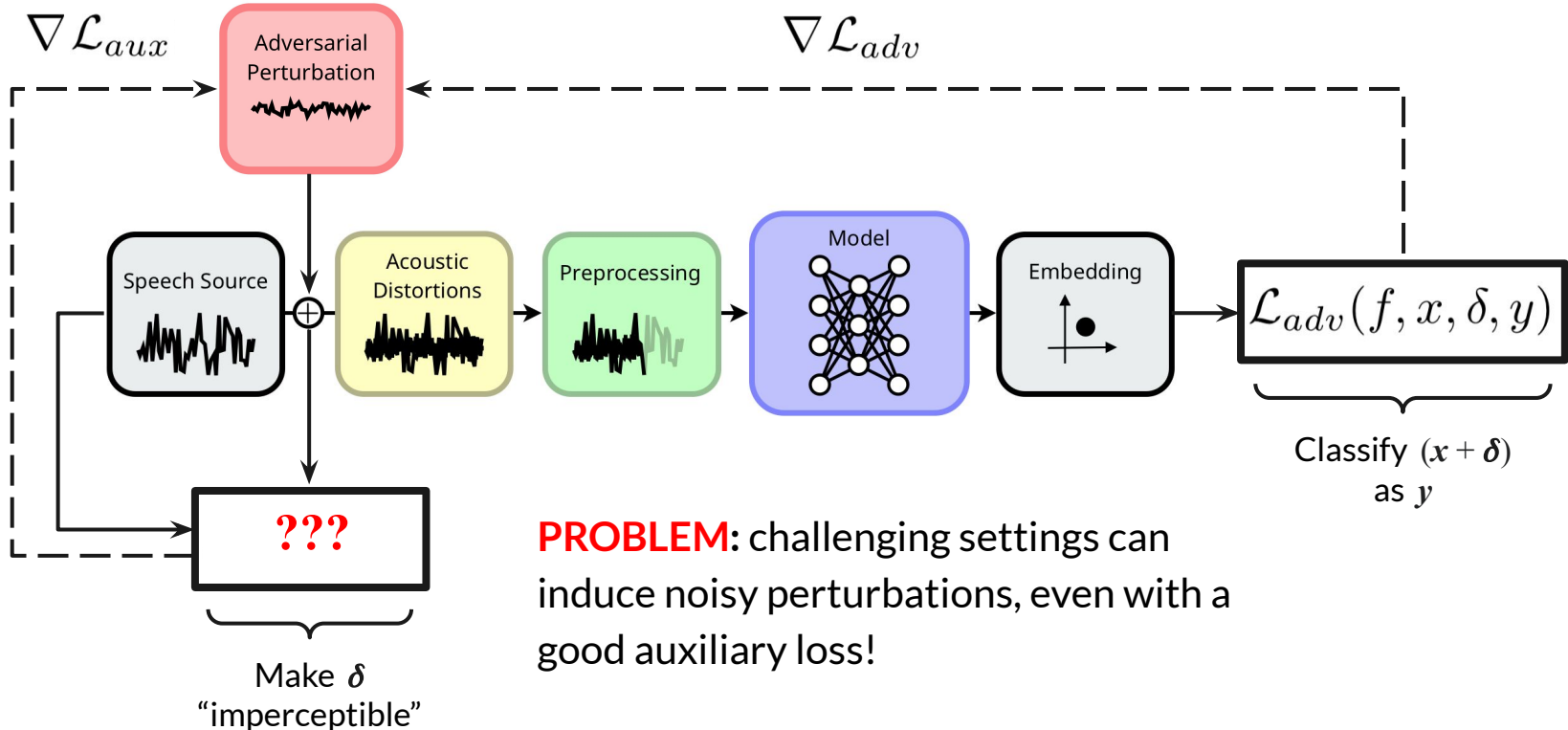


# Let's Attack a Voice Interface: System Design





# Let's Attack a Voice Interface: The Noise Issue



# Let's Attack a Voice Interface: Pick an Attack

---

Qin et al. (2019): speech recognition



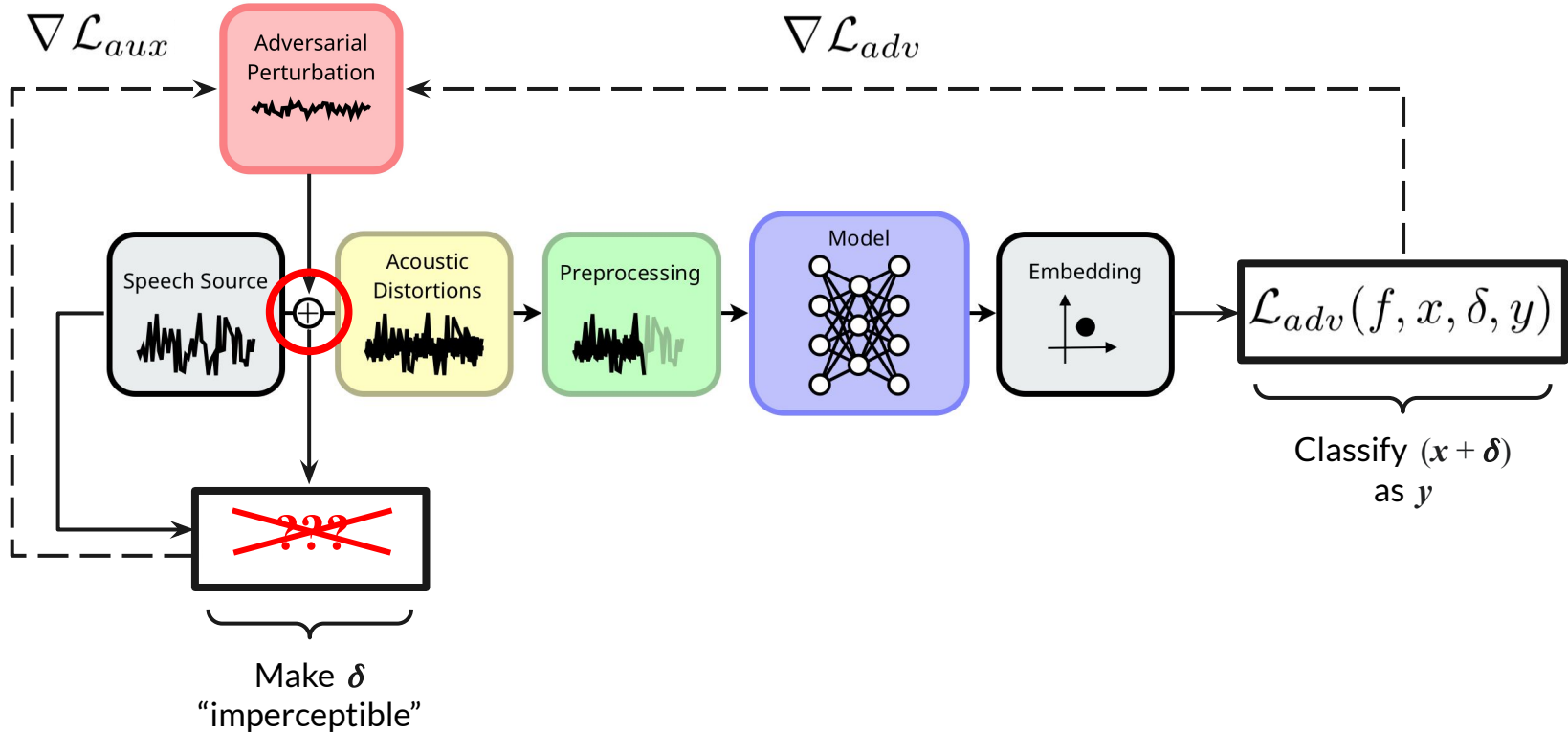
Li et al. (2020): speaker recognition



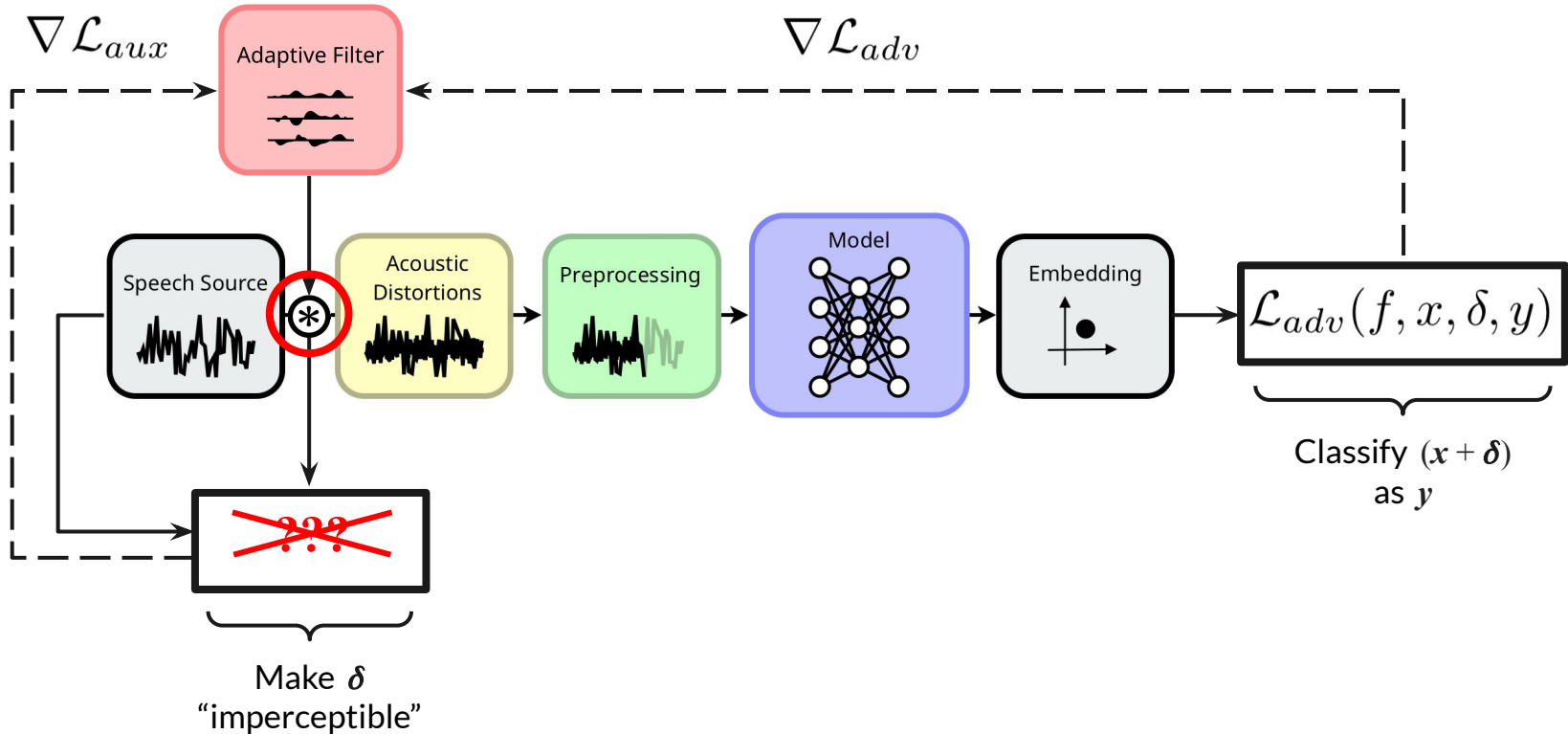
Chen et al. (2020): speech recognition



# Let's Attack a Voice Interface: System Design

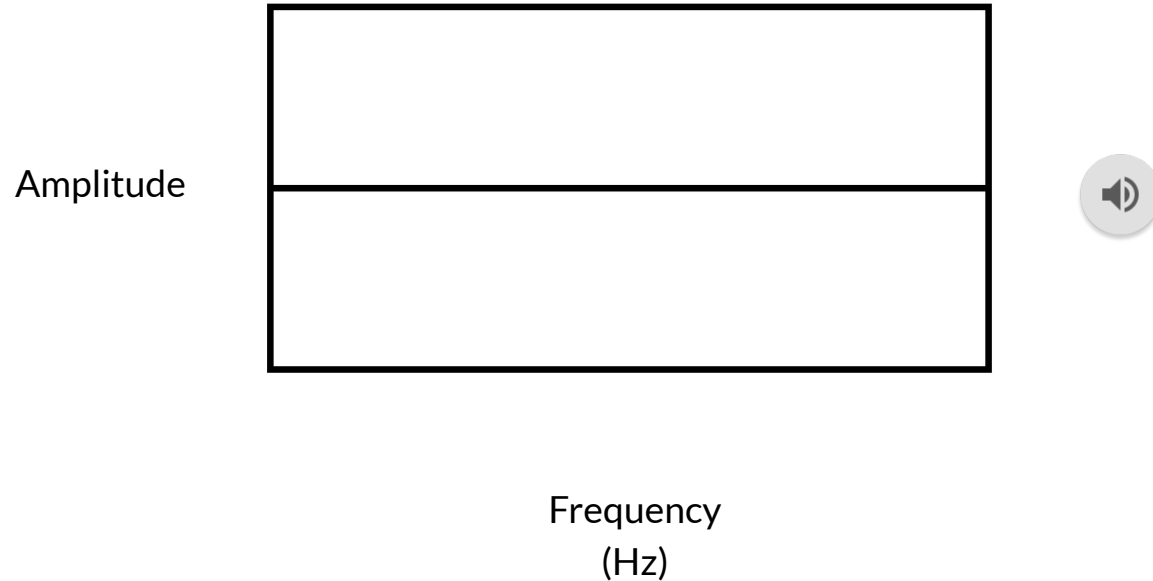


# Let's Attack a Voice Interface: Adaptive Filter Attack



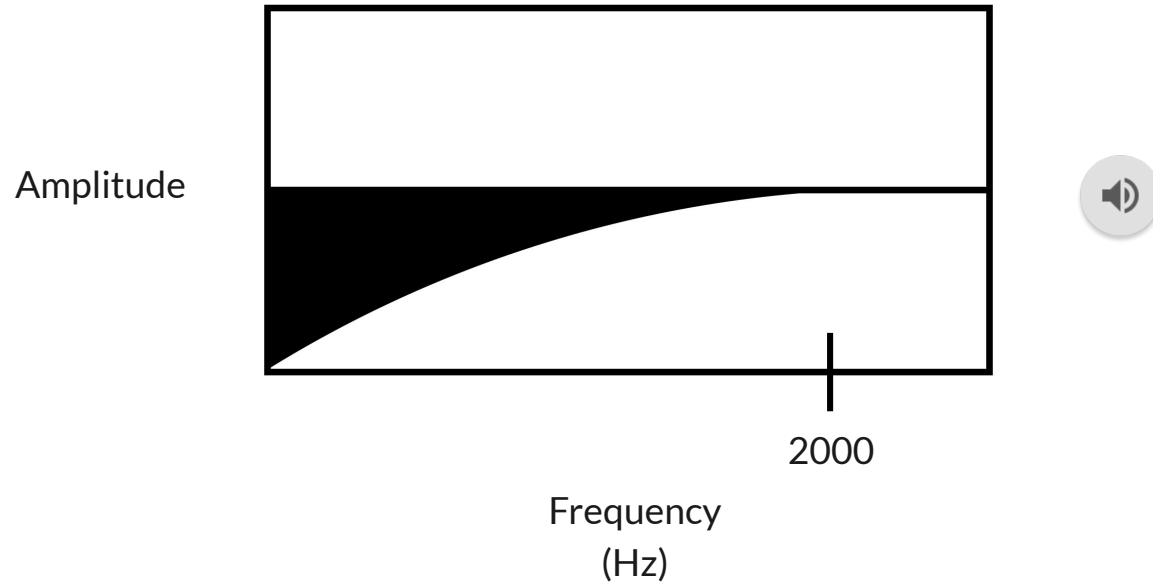
# Adaptive Filters Let Us Shape Frequency Content

---



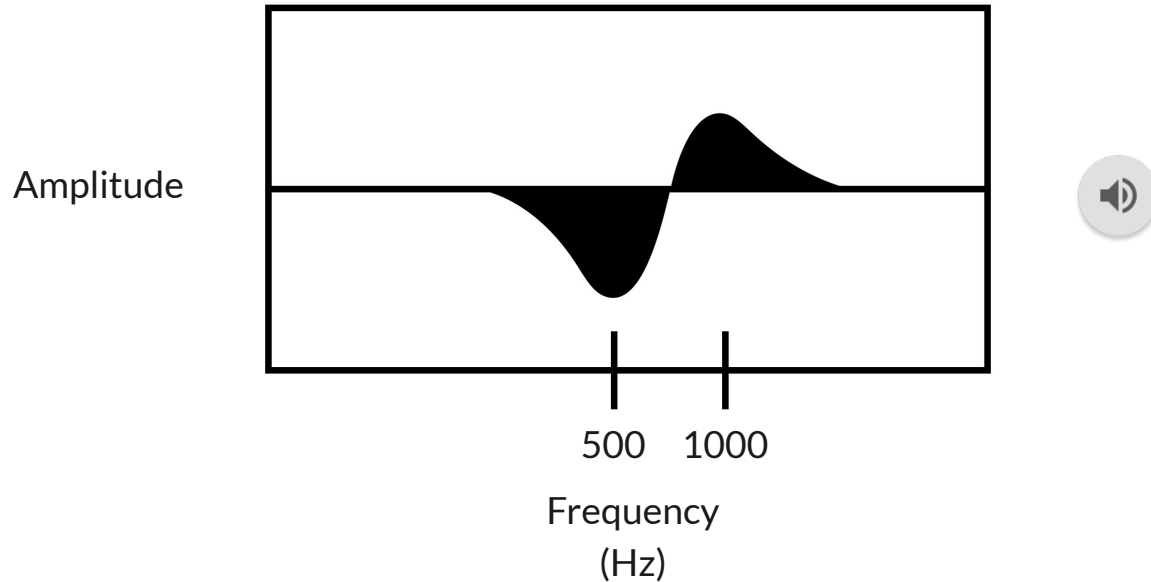
# Adaptive Filters Let Us Shape Frequency Content

---



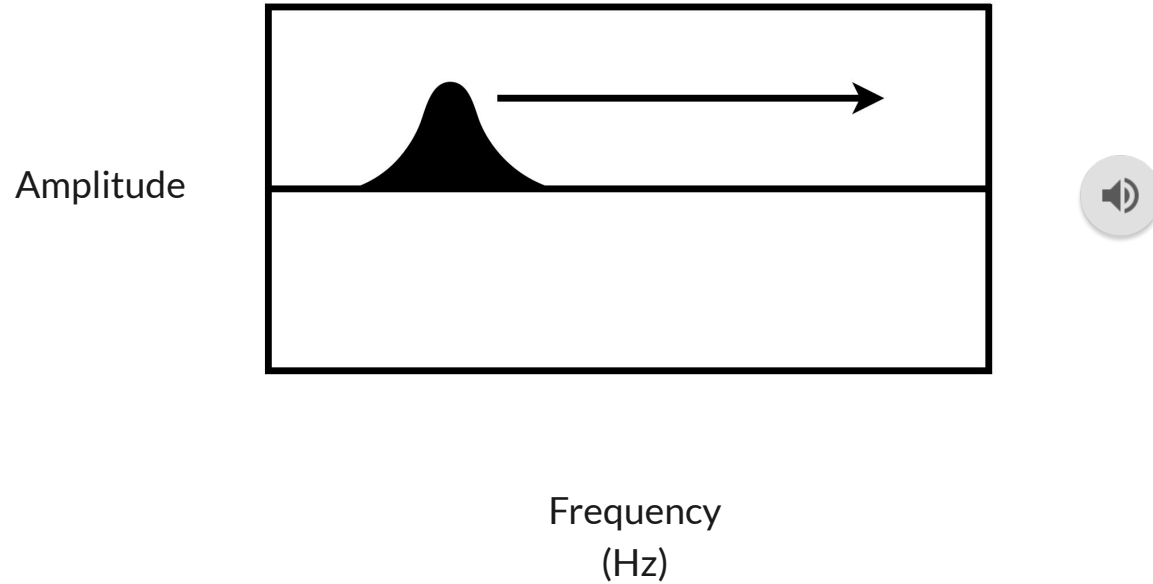
# Adaptive Filters Let Us Shape Frequency Content

---



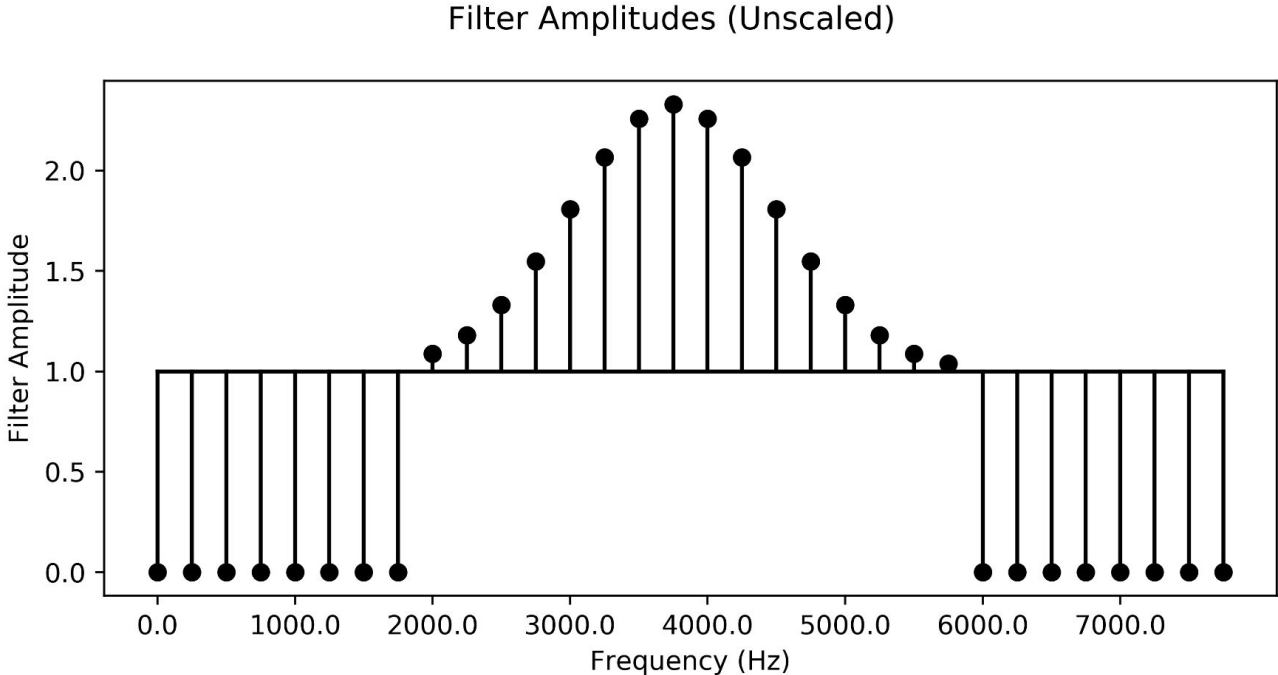
# Adaptive Filters Let Us Shape Frequency Content

---

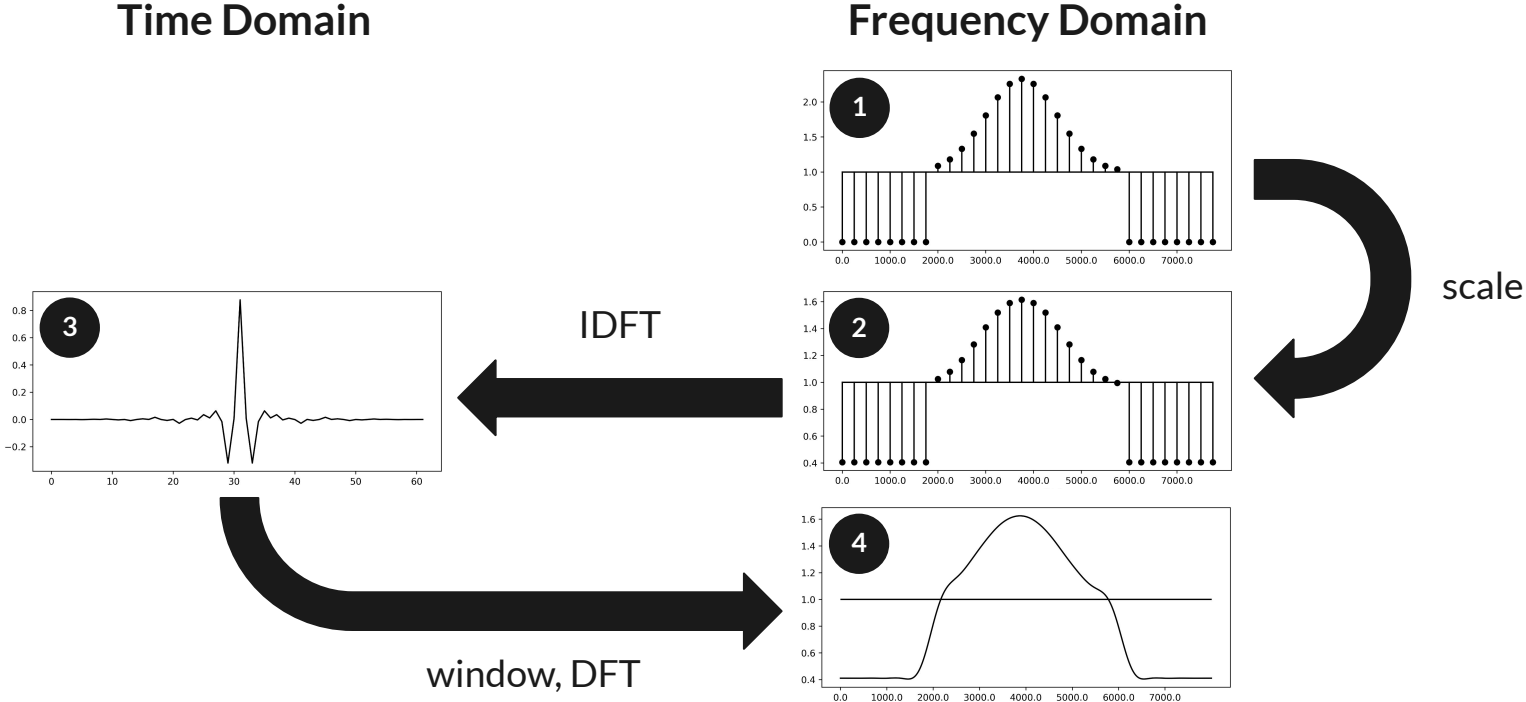




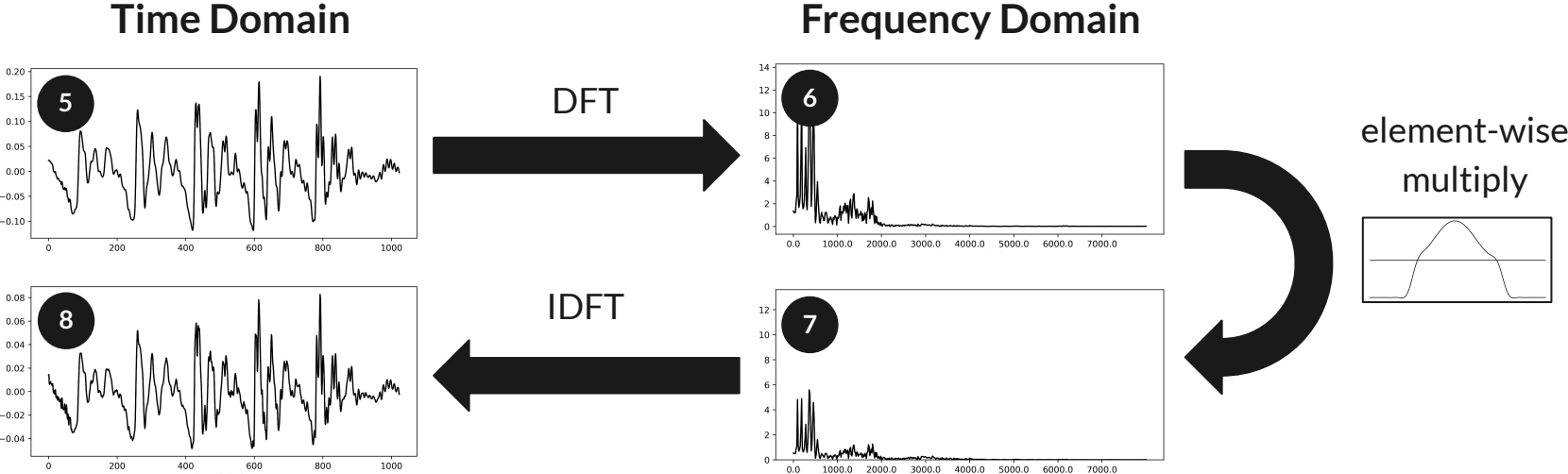
# Adaptive Filters Let Us Shape Frequency Content



# Adaptive Filters Let Us Shape Frequency Content

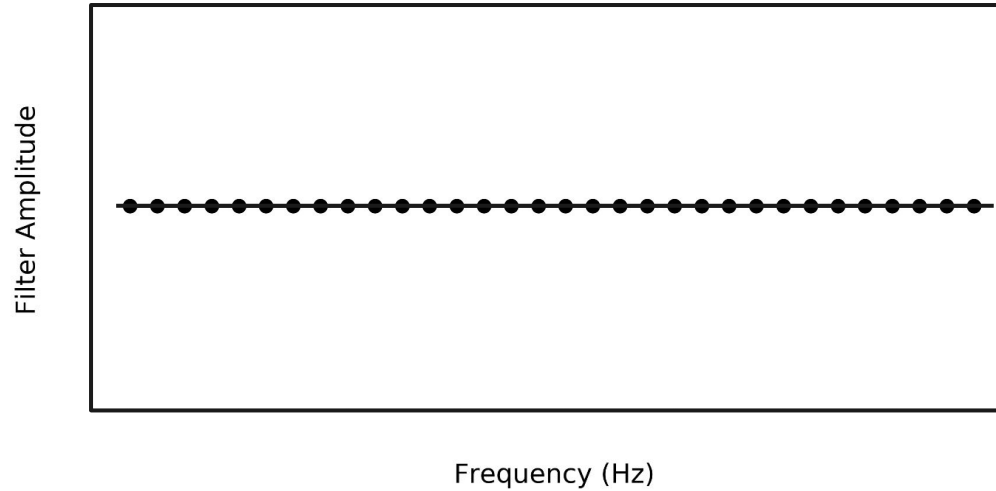


# Adaptive Filters Let Us Shape Frequency Content



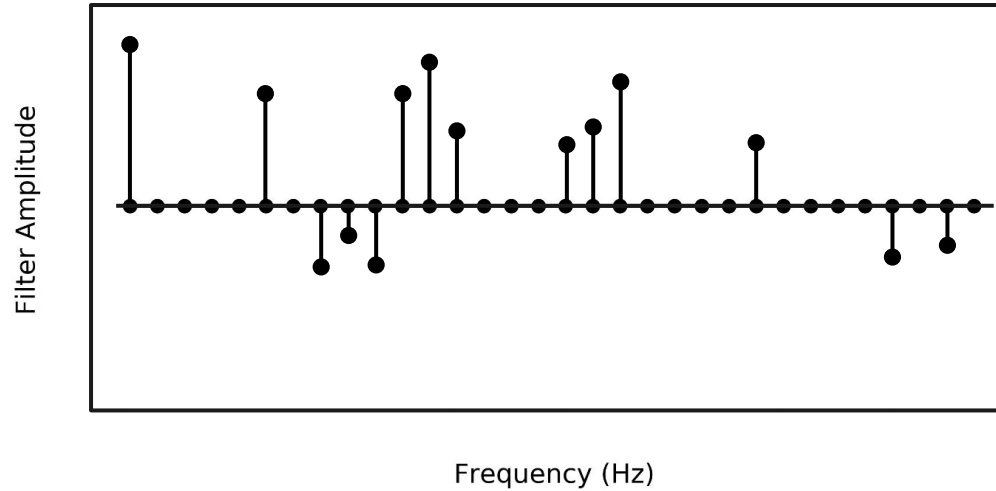
# Adaptive Filters Let Us Shape Frequency Content

---

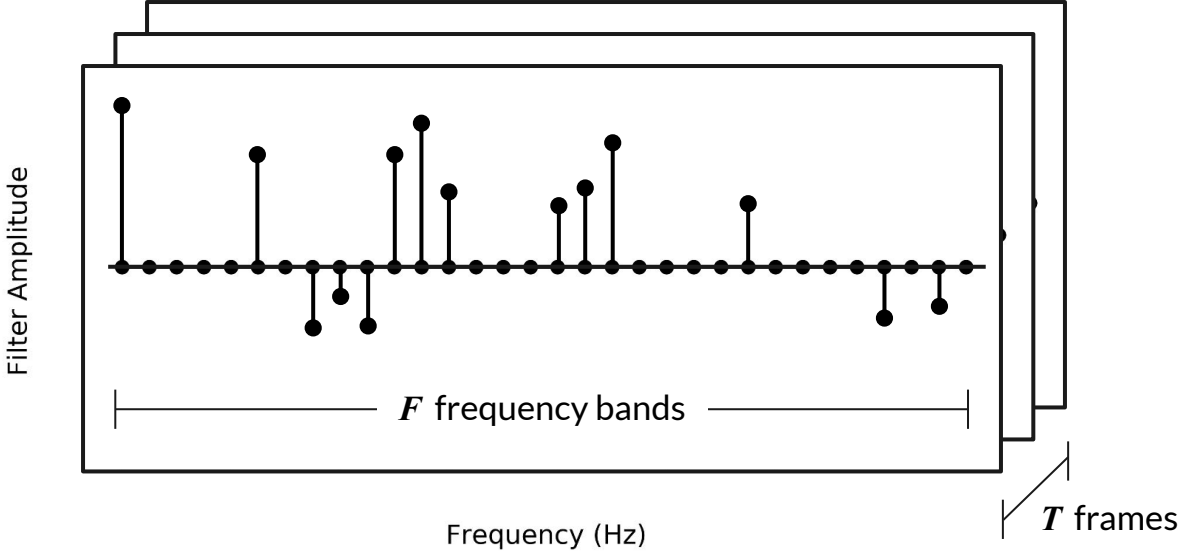


# Adaptive Filters Let Us Shape Frequency Content

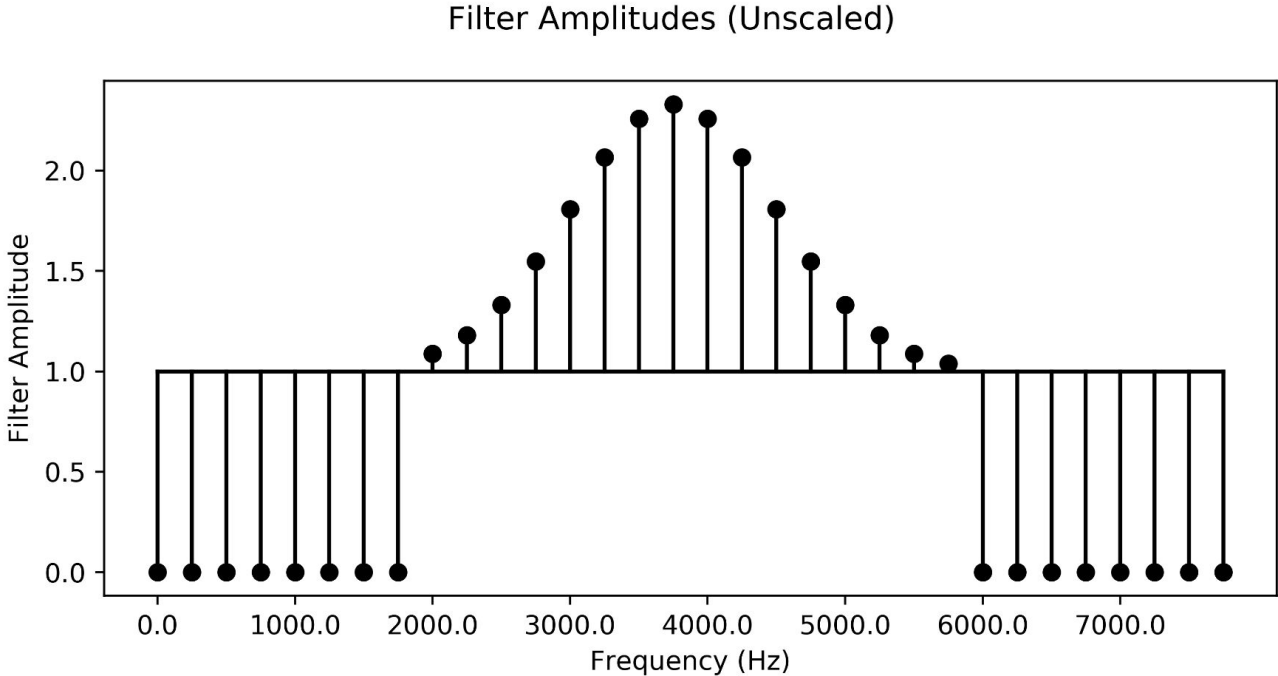
---



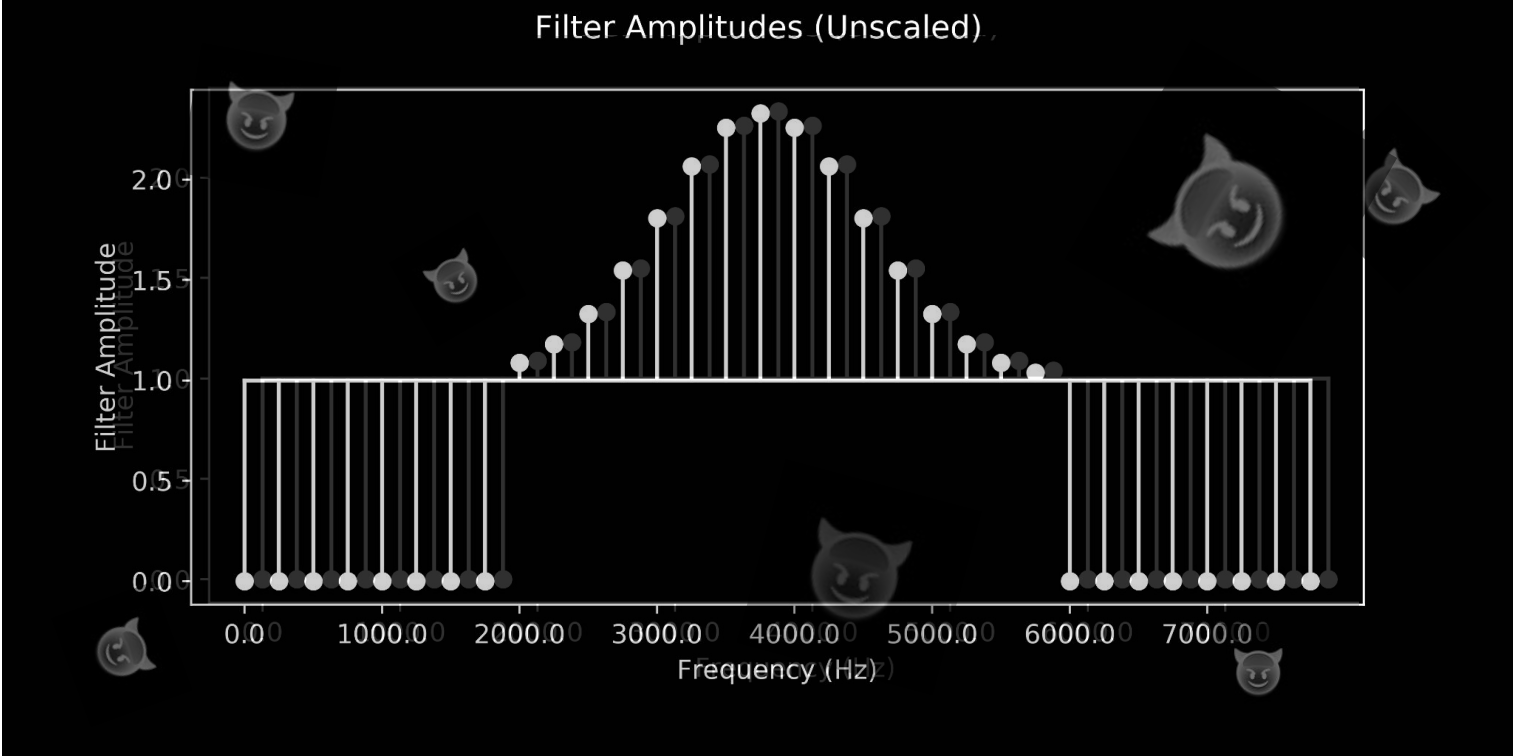
# Adaptive Filters Let Us Shape Frequency Content



# Adaptive Filters Let Us Shape Frequency Content

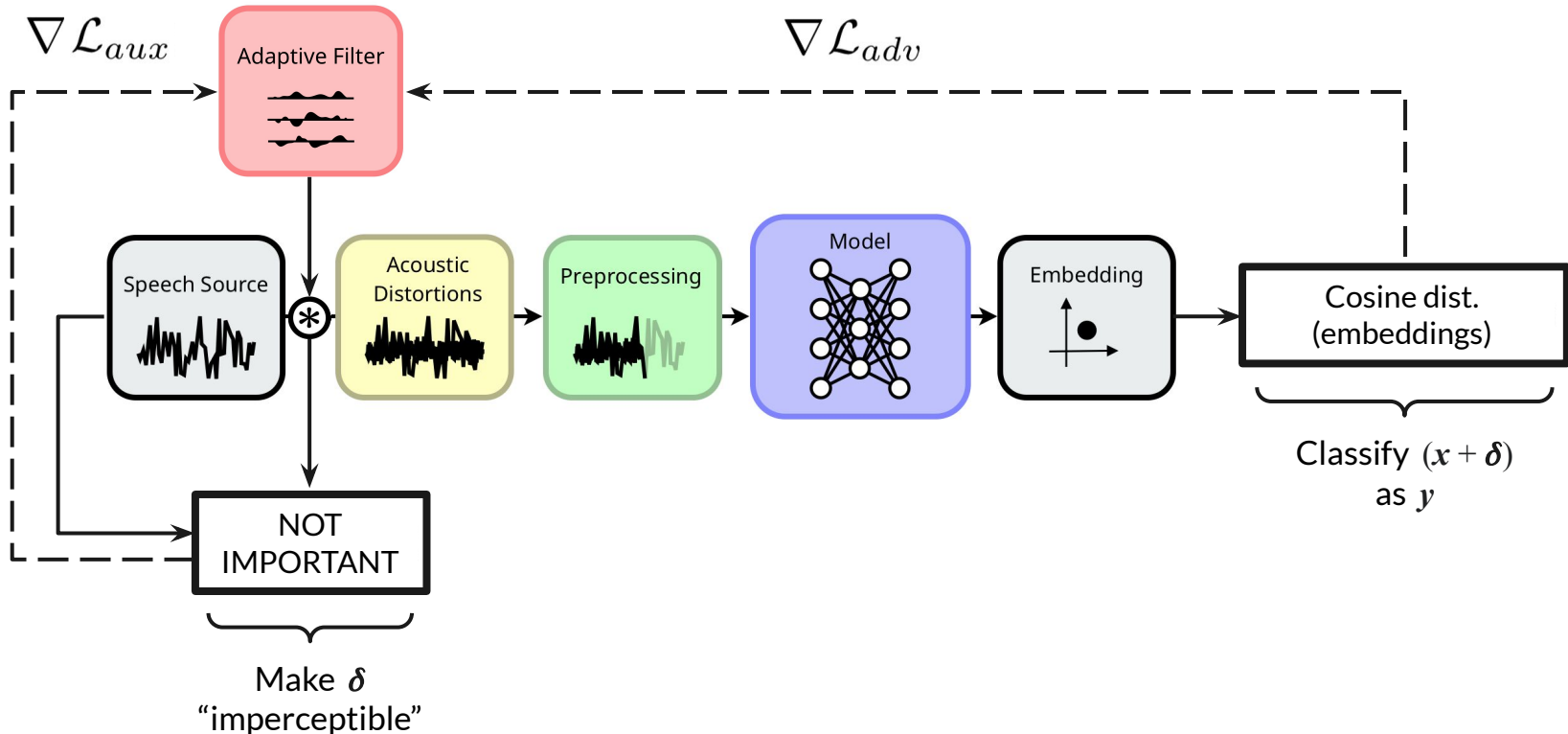


# Adaptive Filters Let Us Shape Frequency Content

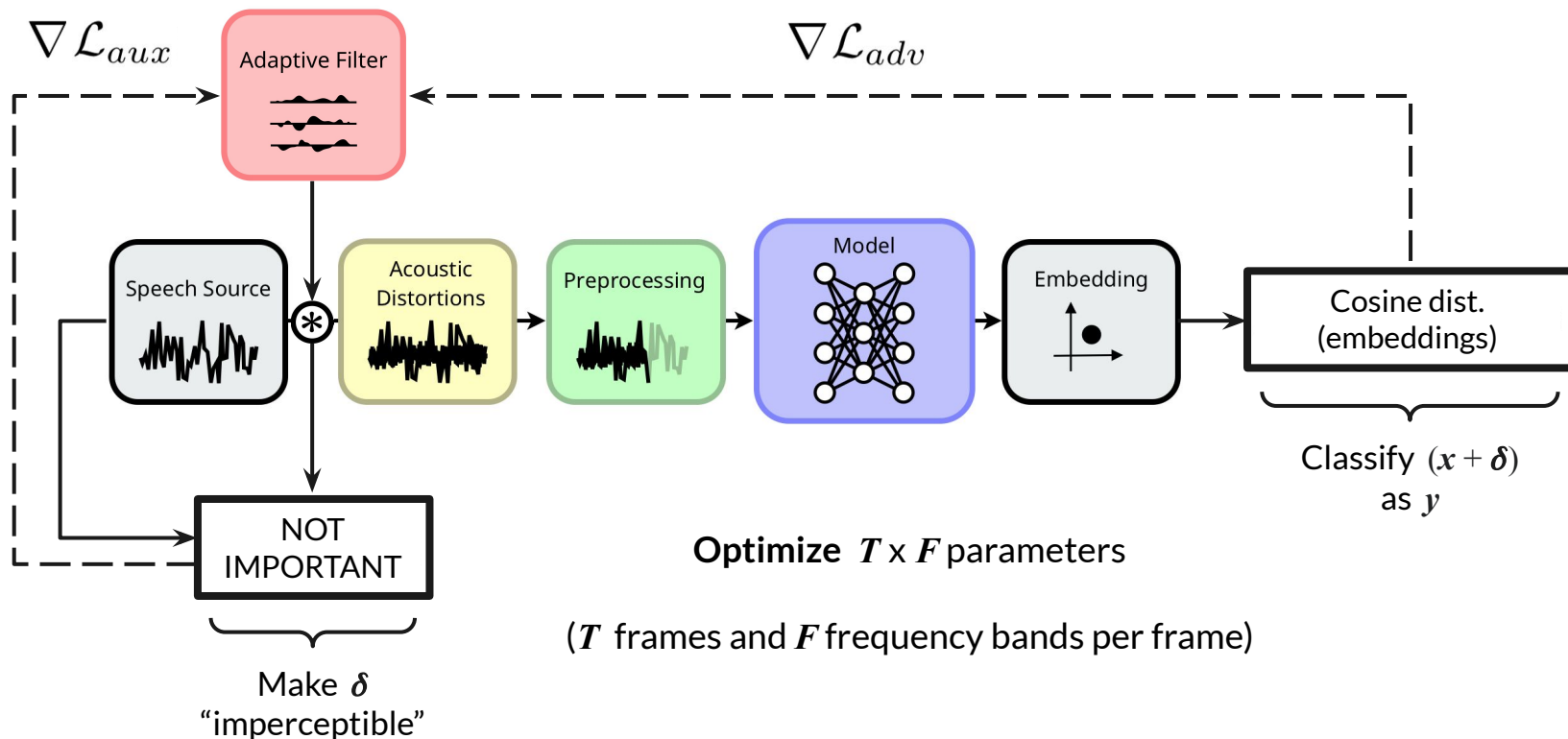




# Let's Attack a Voice Interface: Adaptive Filter Attack

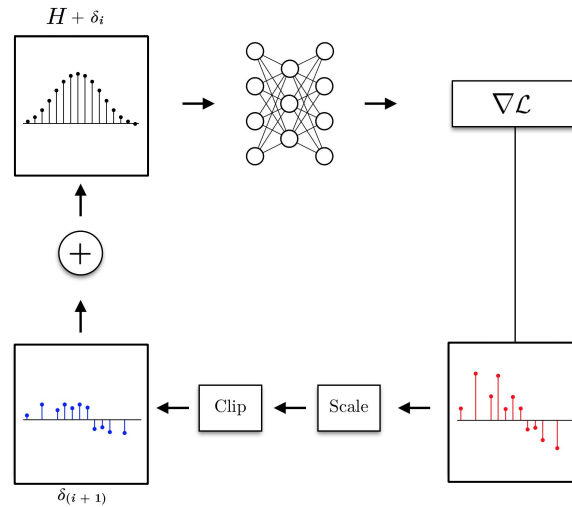


# Let's Attack a Voice Interface: Adaptive Filter Attack



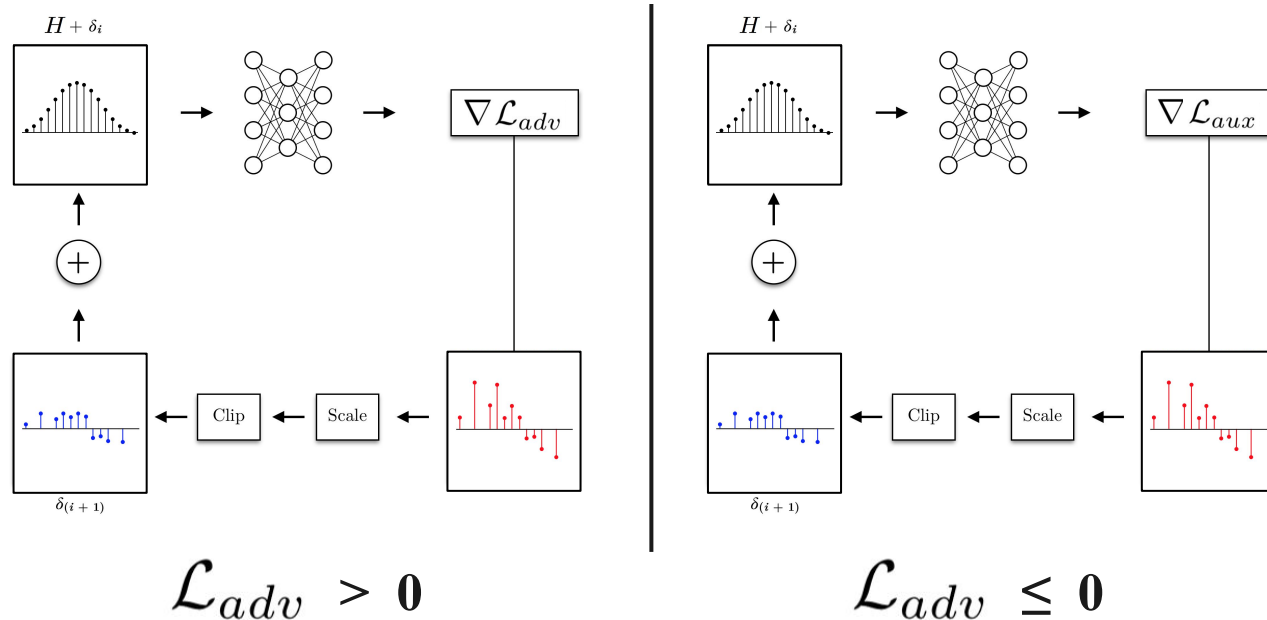
# Let's Attack a Voice Interface: Adaptive Filter Attack

Recall the iterative adversarial optimization procedure we discussed earlier.



# Let's Attack a Voice Interface: Adaptive Filter Attack

*Selective projected gradient descent* (Bryniarski et al. 2021) - break up the updates



# Why Attack with Adaptive Filters?

---

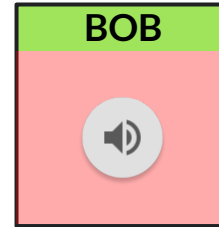
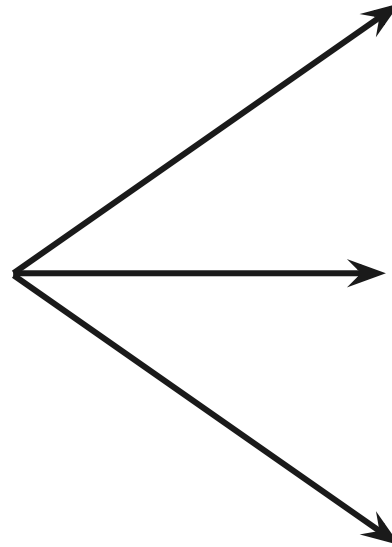
# Why Attack with Adaptive Filters?

---

1. Introducing perturbations at the filter representation, rather than the waveform, avoids noise-like artifacts

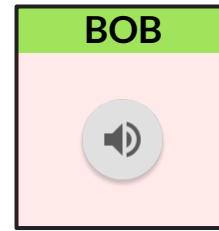
# Why Attack with Adaptive Filters?

---



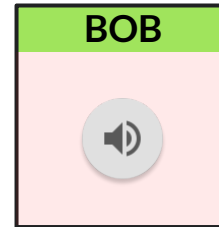
"Generic"

89% effective



Qin et al.\*

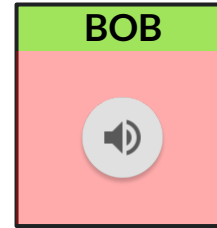
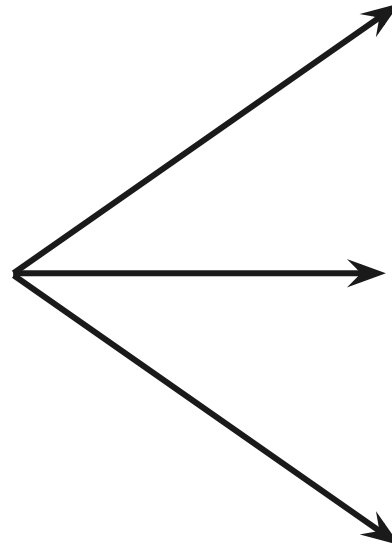
93% effective



**Adaptive Filtering**

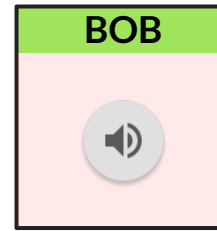
95% effective

# Why Attack with Adaptive Filters?



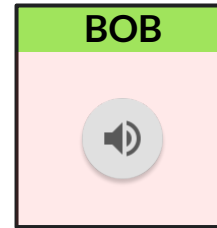
“Generic”

86% effective



Qin et al.\*

90% effective



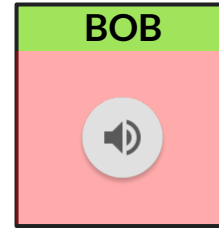
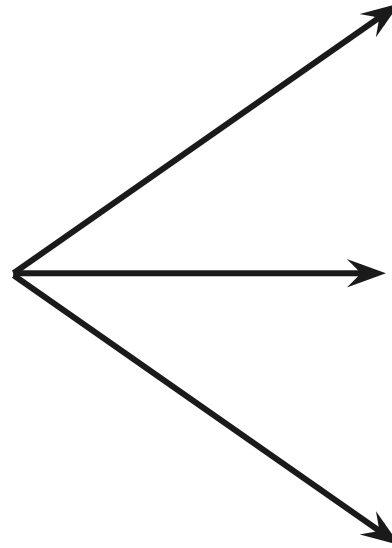
Adaptive Filtering

93% effective

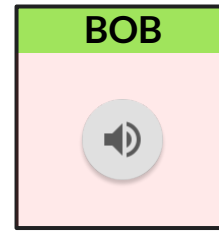
In general, when optimizing for more challenging distortions, attack success rate drops and artifacts become more audible



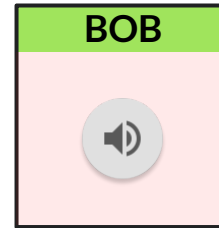
# Why Attack with Adaptive Filters?



“Generic”  
89% effective



Qin et al.\*  
93% effective



**Adaptive Filtering**  
95% effective

**User Study:** if we match effectiveness rates, listeners find our attack less conspicuous than Qin et al.\* by a 2-to-1 margin

# Why Attack with Adaptive Filters?

---

	Waveform $L_\infty$	Waveform $L_2$	<u>Perceptual Study</u> Forced Choice
Qin et al.*	0.08	1.97	34.1%
Adaptive Filtering	0.23	6.59	65.9%

# Why Attack with Adaptive Filters?

---

	Waveform $L_\infty$	Waveform $L_2$	Perceptual Study Forced Choice
Qin et al.*	--	--	--
Adaptive Filtering	<b>2.88x</b>	<b>3.35x</b>	<b>1.93x</b>

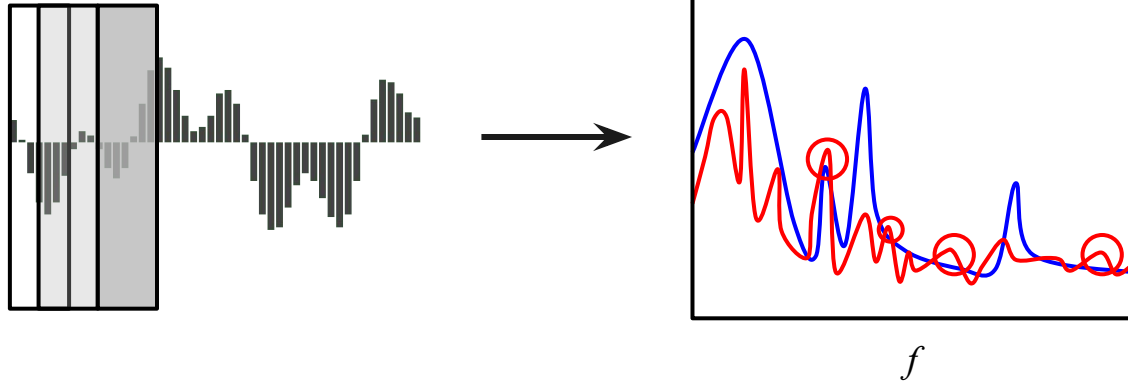
## Why Attack with Adaptive Filters?

---

2. When we use filters, we do not need a complex perceptual loss to produce inconspicuous attacks

# Why Attack with Adaptive Filters?

---



**Two-stage frequency-masking attack:** Qin et al. (2019), Szurley & Kolter (2019), Dörr et al. (2020), Wang et al. (2020)

# Future Directions

---

Other recent works have also begun exploring attacks at representations other than the waveform (e.g. *FoolHD*, *PhaseFool*, *Adversarial Music*)

# Future Directions

---

Other recent works have also begun exploring attacks at representations other than the waveform (e.g. *FoolHD*, *PhaseFool*, *Adversarial Music*)

We plan to explore filter-based attacks against more robust speaker verification pipelines, as well as other speech systems

# Future Directions

---

Other recent works have also begun exploring attacks at representations other than the waveform (e.g. *FoolHD*, *PhaseFool*, *Adversarial Music*)

We plan to explore filter-based attacks against more robust speaker verification pipelines, as well as other speech systems

We also plan to explore the implications of this work for improving the robustness of audio models against large-magnitude frequency-domain perturbations



# Adversarial Attacks in the Audio Domain with Adaptive Filtering

Patrick O'Reilly<sup>1</sup>, Pranjal Awasthi<sup>2</sup>, Aravindan Vijayaraghavan<sup>1</sup>, Bryan Pardo<sup>1</sup>

<https://interactiveaudiolab.github.io/project/audio-adversarial-examples.html>



1. Northwestern University
2. Google Research

**Thanks!**